# LASSO for Public Health Data:

# An Examination of Prevalent Variable Selection Methods and Demonstration of LASSO in R

by

Suzanne M Dufault

A thesis submitted in partial satisfaction

of the requirements for the degree of

Master of Arts

in

Biostatistics

in the Graduate Division of the

University of California, Berkeley

Committee in charge:

Professor Nicholas Jewell
Professor Sandrine Dudoit
Professor Lia Fernald

May 12, 2017

# Contents

# Dedication

I would like to express my gratitude to Dr. Abhijeet Singh, a leader in transparency and collaboration. Thank you for graciously sharing all of the well-organized STATA code necessary for the analysis completed in this paper. I hope the field continues to move forward in this productive manner and will personally do all I can to contribute to this end.

I would also like to thank my committee members. Thank you, Professor Fernald, for the opportunity to explore this topic. Thank you, Professor Dudoit, for further discussion on LASSO and parametric modeling in general. Thank you, Professor Jewell, for ongoing advisement.

Finally, I would like to thank my family. Your unwavering support means everything to me.

# Acknowledgements

The data used in this publication come from Young Lives, a 15-year study of the changing nature of childhood poverty in Ethiopia, India, Peru and Vietnam (YoungLives.org.uk). Young Lives is funded by UK aid from the Department for International Development (DFID). The views expressed here are those of the author. They are not necessarily those of Young Lives, the University of Oxford, DFID or other funders.

# 1 Introduction

Through the development of increased data storage and high-dimensional analytics, researchers are no longer restricted to the small, easily manageable datasets for which traditional statistical analyses were designed. Changes in the way data is not only collected and analyzed, but the sheer amount of data that can be stored and shared, have been heavily felt in public health research. The existence and accessibility of large public health datasets, while providing incredible opportunities for research, have also led to recent uncertainty and controversy over the value of using existing observational data rather than completing a unique study to answer research questions. While this paper does not attempt a deep investigation into the data-sharing debate (Longo and Drazen 2016), a related concern is how researchers are determining which variables to include in an analysis.

This is an issue all researchers must face, but one that is nonetheless critical for those making use of pre-existing databases. A primary example is the groundbreakingly extensive database collected and maintained by the Young Lives Study (YL), from which more than 400 papers have been produced. Young Lives has followed nearly 12,000 children from four countries over a fifteen year span in an effort to better understand the effects of childhood poverty as these children enter adulthood. There are two cohorts of children: the younger cohort born between 2001 and 2002 and the older cohort born between 1994 and 1995. In 2002, 2006, 2009, 2013, and 2016, household and school surveys were administered with three parts: a child questionnaire, a household questionnaire and a community questionnaire. Everything from the height and weight of the children and their caregivers, the household assets, and the religions and languages spoken in the community is recorded (YoungLives.org.uk). Due to the wealth of data collection, papers have been published on poverty, nutrition, education, gender, child protection and more. Further details regarding the cohorts as well as the data collection methods themselves have been well-documented on the YL website.

As mentioned, an immediate concern for the use of such datasets arises: given the wealth of information provided in this database and others like it, how are researchers determining which variables should be included in their analyses? Concerns of *ad hoc* model building, overadjustment and its potential to obscure or bias the effects of the variable of interest, as well as potential losses in precision, prevent researchers from throwing everything that might be useful into analyses (Schisterman et al. 2009). While suggestions have outlined more rigorous and uniform variable selection practices, the extent to which journals and researchers conform to these standards varies considerably. Further, examples of more data-driven methods, such as penalized regression, are almost non-existent in the epidemiologic literature despite enthusiasm for their statistical qualities and improvements on traditional regression techniques (Walter and Tiemeier 2009).

This paper intends to first take a further look at existing variable selection practices in epidemiologic research (Section 2) and provide a brief presentation of the merits and shortcomings of several of the most popular variable selection techniques (Section 3). Then, in Section 4, the LASSO regression and ways in which the method improves upon traditional least squares regression are described. Finally, in Section 5, LASSO is compared to the results from a least squares regression from the published paper "Test Score Gaps between Private and Government Sector Students at School Entry Age in India" (Singh 2014). The LASSO analysis was completed using R and all code used for the analysis in Section 5 has been commented and stored in Appendix **??**.

Considering the growing variety of variable selection techniques, greater documentation of analysis processes should be encouraged if not required by the community. Computational reproducibility, the ability to take an existing dataset and recreate results produced on that same dataset, seems straightforward. Yet, a culture of sharing methods and analysis techniques does not appear to be prominent in the field. To take direct action towards greater transparency, this entire report has been written using "knitr" in R Studio. All source files are currently available upon request and, as the data used here is the property of the UK Data Service, actions are being taken to create a simulated dataset with which examples from the code in this report can be run.

# 2 Literature Review

## 2.1 Motivation

The 2009 paper *Variable Selection: Current Practice in Epidemiological Studies* (Walter and Tiemeier 2009) provides an analysis of 300 articles found in four of the most popular epidemiological journals: *American Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology*, and *International Journal of Epidemiology*. Recognizing that variable selection is one of the most controversial parts of research, the authors were interested in gaining a sense of what methods were being used and accepted and whether methods such as shrinkage and penalization, referred to as "modern methods" by the authors, had been incorporated into research.

While their investigation produced a wealth of insight into the current and acceptable practices of modern research, a few of their findings merit particular attention. The authors found that 35% of all articles surveyed made no mention of their variable selection techniques. This proportion includes papers that stated their variables were selected via *a priori* knowledge or information but provided no explicit citations to the source of the knowledge. While this lack of information is concerning for the sake of reproducibility and transparency, it is not a sign that the research in these articles is inherently flawed. Instead, it demonstrates the importance of reflecting on why journals do not prioritize and require richer detail for this consequential aspect of the analysis.

An additional 27.667% of articles sampled completed variable selection based on previous literature and cited prior knowledge. While this has generally been recognized as a reasonable practice (Sauer et al. 2013), it certainly has its flaws. In a theoretical sense, concerns about scientific findings that may actually be null should give at least slight pause to the unquestioning inclusion of such variables (Ioannadis 2005). Practically speaking, tracing these citations back to the original study in which the variable was proven to be associated with a particular outcome of interest often requires a search through several generations of publications before the original source is found. Further, even rigorously selecting variables based on prior knowledge may still leave a researcher with a large and unwieldy subset, requiring an additional method of screening for variable importance.

Finally, not a single article in the sampled set of 300 was found to make use of shrinkage or penalization methods. While I will explore further the attributes of such a method, this was quite a shock to the original authors given the support such methods had received from methodologists (Greenland 2008; Sauer et al. 2013). Pairing this information with the finding that 34% of all papers still used the highly criticized methods of step-wise selection and change-in-estimate, it is clear that methods that have dominated the literature in the past continue to dominate in the present, regardless of their catalogued benefits and failures.

## 2.2 Recent Review of Young Lives Journal Articles

In April of 2016, I completed a literature search on PubMed and Google Scholar searching for a more recent analysis of method prevalence in epidemiologic literature. Given that the 2009 paper is nearly a decade old, it was hypothesized that there may have been a shift towards shrinkage and penalization methods given the increasing ease of implementation provided by user friendly statistical packages. I was unable to find a comparable comprehensive review.

In light of this, a considerably smaller review was completed, narrowed to the scope of the journal articles produced regarding the YL study. By June 2016, there were 145 published journal articles referenced on the YL website. Conditioning the sample based on journal articles with any transformation of the cognitive measures of the Peabody Picture Vocabulary Test (PPVT) or Cognitive Development Assessment (CDA) as a primary outcome of interest, returned a sample of 19 eligible papers. These outcomes were chosen to ensure similarity in motivation as well as restrict the sample size of eligible papers. The cognitive measures are well described in other resources (Singh 2014). Table 1 summarizes the variable selection methods found in these 19 papers.

| **Selection Technique** | **n** | **Percent** |
|---|---|---|
| Prior knowledge | 9 | 47.37% |
| Effect estimate change | 2 | 10.53% |
| Stepwise selection | 0 | 0 |
| Modern methods (shrinkage, penalized regression) | 0 | 0 |
| Other (e.g., principal components, propensity scores) | 3 | 15.79% |
| Not described | 5 | 26.32% |
| Total | 19 | 100% |

Table 1: A summary of variable selection techniques used in Young Lives journal articles regarding quantitative measure of cognitive outcomes as the outcome of interest.

Similar to the results of the 2009 paper, Table 1 shows approximately 47% of papers in this sample used prior knowledge for variable selection, 26% of papers did not describe the variable selection process at all and none of the sampled papers used shrinkage or penalization methods for variable selection. Additionally, there was considerable homogeneity in analysis techniques. 73% of papers used un-penalized parametric regression models to estimate their parameters of interest. While the prevalence of the technique does not confirm or deny its efficiency in answering research questions, it is clear that researchers and journals are generally comfortable using and interpreting regression models. As such, the rest of this paper will focus on variable selection within the framework of linear regression modeling. The value of the shared understanding of regression will be leveraged in Section 4 when introducing LASSO.

# 3 Brief Review of Existing Methods for Variable Selection

As is evident in Table 1, there are a number of ways to complete variable selection. Before describing several of these methods, it is important to consider the motivation for engaging in variable selection. Consider the following scenario, common in public health research. A dataset contains measurements on $n$ different individuals and $P$ total covariates per individual. One of the most common objectives is then to explore the relationship of the explanatory covariates $X$ with the outcome $Y$, statistically expressed as $E[Y|X]$. As discussed earlier, this relationship is most commonly explored by the use of a linear regression of the observed outcomes $Y_n \in \mathbb{R}^{n\times 1}$ on the observed covariates $X_n \in \mathbb{R}^{n\times P}$ by estimating the coefficients $\beta$.

One of the criteria commonly used for selecting the "best" estimator of a parameter is the mean squared error (MSE) of the estimator. Expressed in Equation 1, where, in the context of a linear regression, $\hat{\beta}$ is the estimator and $\beta$ is the true parameter value, the MSE relies on the bias and variance of the estimator $\hat{\beta}$. In order to minimize the MSE, one then tries to find estimators that have minimal bias and minimal variance.

$$MSE_{\hat{\beta}} = E_{\beta}[(\hat{\beta} - \beta)^2] = Var_{\beta}(\hat{\beta}) + (Bias_{\beta}(\hat{\beta}))^2 \tag{1}$$

Practically speaking, this bias-variance tradeoff is the constant battle of finding estimators that approximate the truth well, and as such fit the data well, but are also very stable. Stability can be visualized as the smoothness of a regression line. For example, if an outcome $Y$ is modeled as $E[Y|X] = a$, where $a$ is the mean of $Y$, then for any $X = x$, the estimate of $Y$ will always be its marginal mean. This is a very stable estimator: it does not change at all regardless of the size of change in $X = x$. It is also highly biased if $E[Y|X]$ does depend on the realization of $X$. In this example, the tradeoff has been made to find an estimator that has low variance but the potential for high bias. If a model makes use of too many covariates, the estimator may be less biased, but highly variable. Further, the more terms included in a linear regression model the greater the concern of overfitting, or modeling noise that is specific to the dataset at hand rather than the true underlying signal. More moderate tradeoffs such as a small increase in bias can lead to a large decrease in variance, resulting in a smaller MSE overall. This is the case with LASSO, which will be discussed further in Section 4.

Linear models run into problems when $P$ is much larger than $n$, when $P$ is equal to $n$, or when $P$ is slightly less than $n$. In each of these settings, researchers may be interested in selecting an "optimal" subset of covariates $p \in P$ to use in the linear regression instead of using the entire original set of $P$ covariates. With respect to intepretability, models that make use of fewer variables are typically easier to understand and the variables are conceptually simpler to map with respect to each other than a large number of variables with complicated relationships. When researchers have *a priori* knowledge as to the potential relationships of particular variables (e.g. an exposure of interest and well-defined potential confounders) a regression model can be used to test these *a priori* hypotheses. Regression with fewer variables can also be desirable for more exploratory analyses, as once again, interpretability is generally a primary objective.

To add to the difficulty of including enough covariates so as to accurately represent the complexity of the world while simultaneously finding low variance estimators, a great number of methods and criteria exist for selecting an "optimal" subset of covariates. While efforts have been made to summarize the existing literature into a series of best practices (Sauer et al. 2013), consistent application or even agreement on these practices have yet to have been adopted by the field at large. In the following paragraphs, I attempt to briefly catalogue the most common statistical techniques for variable selection in the context of linear regression models as well as provide an overview of their limitations.

## 3.1 Subset and Stepwise Selection

The *best subset* method aims to determine the optimal subset of covariates by an exhaustive search through the $P$ possible variables. In this setting, every possible subset of variables, of which there are $2^P$, is tested and optimality is determined by a global certain criteria (Christensen 2011, pp. 381-385). A variant of the method minimizes the sum of squared errors by testing every possible subset of variables of a predetermined subset size $p \in P$. In either setting, this is a computationally expensive method in that every variable is included and excluded in every possible combination. Further, there are considerable multiple testing issues that arise, worsened by the implementation of criteria that was originally meant to test hypotheses of effect size not model selection (Dziak et al. 2005).

*Stepwise selection*, typically via forward or backward steps, significantly shrinks the number of possible subsets one must consider in order to find the optimal subset and essentially functions as an approximation of the *best subset* method. For *forward selection*, the intercept is first fit. In a sense, this is the first subset and comparing any other model against this first subset investigates whether an additional variable makes a significant contribution to the fit of the model. A few of the ways in which contributions can be considered significant includes observed increases in the model $R^2$ via an $F$ statistic comparing the sum of squared errors from the smaller subset with those from the larger subset, via $T$ tests regarding whether the additional coefficients on a larger model are significantly different from the null, or by the largest increase in absolute partial correlation (Christensen 2011, pp. 385). This process is continued, adding one variable at a time to the model (always that which helps to best improve on the decided criteria) until adding an additional variable does not significantly improve the fit. *Backward selection* follows the same path, but starting from a full model with all covariates included and removing covariates until a particular stopping criterion has been met.

These stepwise methods improve on *best subset* selection by decreasing the number of possible subsets, but still fall prey to faults in stopping criteria and multiple testing, among other things. Many stopping critera focus primarily on local performance rather than global performance. Hence, local stopping criteria may lead an individual to limit variables to a locally most efficient though not globally most efficient model. Local stopping criteria also have difficulty when detecting small but important improvements and can prematurely stop model building when the gains to adding additional coefficients plateaus prematurely. Finally, while the gains in computational efficiency were at one time appealing, new software have made the reliance on step-wise variable selection to find a decent model, though almost certainly sub-optimal, unnecessary and undesirable. The faults of this method have been well documented in text books (Hastie et al. 2008; Christensen 2011) and research papers alike (Greenland 2008).

## 3.2 Effect Estimate Change

*Effect estimate change* has proven via simulation studies to be quite effective at identifying confounders, which are often the variables one hopes to control for in models exploring the effect of a particular exposure or set of exposures. Consider the model in Equation 2, where $Y$ is the outcome of interest, $X$ the exposure(s) of interest and $W$ the potential confounder(s). The general premise is that a confounder $W$ obscures the relationship between the exposure of interest, $X$, and the outcome, $Y$.

$$E[Y|X,W] = \alpha + \beta X + \gamma W \tag{2}$$

For the change in estimate criteria, one first fits the unadjusted model, Equation 3, which does not include the confounders $W$. Then after fitting the adjusted model, Equation 2, a comparison of the coefficient estimate from the unadjusted model $\hat{\kappa}$ is compared to the coefficient estimate $\hat{\beta}$ from the adjusted model.

$$E[Y|X,W] = \alpha + \kappa X \tag{3}$$

Comparing the effect estimate, $\hat{\kappa}$, in a model without control for the hypothetical confounder to that which includes the hypothetical confounder$\hat{\beta}$, a 10% change in the estimated effect as measured by $(\hat{\kappa} - \hat{\beta})/\hat{\kappa}$, suggests the variable $W$ is a confounder.

This premise, while attractive, presents another set of concerns. First, the subset of variables considered as potential confounders must be identified. Recommendations have been made to follow the causal mapping formalized by Robins in 1987. Once potential confounders have been identified, one must be wary of controlling for too many confounders when the sample size is not infinite. Regardless of the confounders' theoretical importance, including too many confounders can result in overstratification of the data, unwieldy sparsity in the covariate distribution and instability in effect estimates. Second, there must also be in place *a priori* criteria as to how large of a change in effect must be observed to consider the added variable to be a confounder (Robins and Greenland 1986). Historically, a 10% cutoff has been observed, whereby effect estimates that change by 10% upon the inclusion or removal of a hypothetical confounder are said to display evidence of confounding. However, a recent paper published in the *Journal of Epidemiology* demonstrated that 10% may not always be appropriate. In this particular paper, via simulation and utilization of the NHANES dataset, Lee demonstrated that it would be better to first examine the change in estimate for the standardized exposure and standardized outcome with the inclusion of a random variable simulated from a standard normal distribution. Then, variables that induce a change in estimate greater than the 95th percentile of the previous model including the standard random variable would be considered confounders (Lee 2014). In a sense, this compares the observed change in estimate between the exposure and confounder of interest with the estimated change in estimate associated with the inclusion of a completely independent random variable. While reasonably simple, the 10% cutoff still seems to prevail in research without the use of suggested corrective processes.

## 3.3 Propensity Scores

*Propensity score methods*, proposed by Rosenbaum and Rubin (1983), aim for similar goals as those in Subsection 3.2. Used to control for confounders in non-experimental studies where there is an exposure (or set of exposures) of interest $X$, *propensity score methods* examine the probability of a certain exposure $X$ given a set of variables $W$. This is often completed via a logistic model, as in Equation 4, but can be done non-parametrically as well.

$$g(W) = Pr(X = 1|W) = \frac{e^{\alpha+\beta W}}{1 + e^{\alpha+\beta W}} \tag{4}$$

Unlike Equation 2, where $W$ was assumed to be a set of confounders, the propensity score method considers a set of baseline covariates $W$ and assumes that controlling for propensity score considering $W$ is sufficient to break most confounding. This is implemented in Equation 5. Mathematically, this is an independence assumption: assume an individual's exposure, once their background has been controlled for, arose independently of their future outcome, $Y \perp X|g(W)$ Mimicking randomized control trials, observations with similar propensity scores are compared to each other in order to understand the relationship of the exposure $X$ and outcome $Y$ in the assumed absense of external confounding (Rosenbaum and Rubin 1983).

$$E[Y|X, W] = \alpha + \beta X + \gamma g(W) \tag{5}$$

This technique has proven useful when the outcome of interest is rare, the exposure is common and other methods of variable selection would lead to unmanageable sparsity. However, when the outcome is not rare and other methods of variable selection may be appropriate, it can be difficult to determine the benefits of continued use of controlling for propensity scores. A 2006 paper reviewed the growing use of *propensity score methods* in these settings and found no empirical evidence of improvements in performance when compared to other appropriate confounder identification and control methods (Stürmer 2006).

# 4 LASSO: Improvements on Least Squares

## 4.1 Least Squares: Definition

Once the subset of $p$ observed variables of interest have been selected and the $n$ observations corresponding to these variables have been organized into a matrix $X_n \in \mathbb{R}^{n\times p}$, linear regression is one of the most popular techniques for estimating the parameter(s) of interest, typically the association of the outcome $Y$ with the variables in $X$. Simple linear regression refers to models of the form expressed in Equation 6, where the outcome $Y$ can be expressed as some linear combination of the $p$ observed variables in $X$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon = X\beta + \epsilon \tag{6}$$

Estimation of the coefficients $\beta$ in this model can take many forms, though the most common is through least squares minimization. Estimation of $\beta$ via least squares requires the minimization of the quantity presented in Equation 7.

$$\hat{\beta}_{LS} = \arg\min_{\beta\in\mathbb{R}^p} ||Y_n - X_n\beta||_2^2 \tag{7}$$

Solving the minimization problem in Equation 7 results in the estimator, defined in matrix form, in Equation 8.

$$\hat{\beta}_{LS} = (X_n^T X_n)^{-1} X_n^T Y_n \tag{8}$$

As mentioned in Section 3, there are many criteria by which to evaluate the optimality of an estimator. Though many formulations of the least squares criteria can be found in textbooks (Wasserman 2004; Christensen 2011) and across varying fields (Angrist and Pischke 2008), the results are the same. Least squares estimation provides a reliable, and by some criteria optimal, estimate $\hat{\beta}_{LS}$ of $\beta$ when a series of assumptions are met. First, $\hat{\beta}_{LS}$ is unbiased assuming $E[Y|X] = X\beta$ is linear. The second assumption is that the variance of $Y$ at each level of $X$ is constant $\sigma^2$, alternatively referred to as homoskedasticity. These two assumptions are all that is necessary to ascertain that the $\hat{\beta}_{LS}$ is the best linear unbiased estimator (BLUE), a desirable optimality among all linear unbiased estimators. A final assumption that $Y$ is normally distributed with distribution $N(X\beta, \sigma^2)$, connects the least squares estimates $\hat{\beta}_{LS}$ to those found via maximum likelihood estimation $\hat{\beta}_{MLE}$. The efficient estimation of the $\hat{\beta}$s relies on the independent and identical generation of the pairs $(X, Y)$ or the conditional independence of $Y$ given $X$ (Hastie et al. 2008).

When all of these assumptions are met and a least squares regression model is used for data analysis, considerable resources have demonstrated the ease of understanding the coefficients $\hat{\beta}_{LS}$. In practice, the implementation of the method in STATA, R and SAS packages that are very user friendly has further removed any computational barrier to the method.

With respect to the ease of interpretation, the visualization of linear relationships when considering a change in one of the variables $X_k$ for $k = 1, 2, \ldots$ is very simple. Finally, given the distributional assumptions, confidence intervals and p-values may be estimated for the coefficients $\hat{\beta}_{LS}$. While the unreliability and flaws of p-values to determine statistically significant discoveries have been well documented, these remain a primary source of interpretation and discussion for statistical analyses (Wasserstein and Lazar 2016).

## 4.2 Least Squares: Concerns for Public Health Data

The benefits of least squares estimators are nearly indisputable when all the necessary conditions are met. However, in public health data, this may rarely be the case. First, reliable estimation of $\hat{\beta}_{LS}$ relies on the invertibility (i.e. nonsingularity) of the matrix $X_n^T X_n$. This matrix becomes singular if the columns of $X_n$, i.e. the variables, are not linearly independent, which results in a failed attempt at a unique solution. In public health data, high collinearity is a considerable problem. For example, in some regions age may be highly collinear with years of education for a student population. Given that some students will start school earlier or later than their peers, the two variables summarizing age and years of school, will not be perfectly linearly dependent, but will nonetheless display a high level of collinearity. This high collinearity produces a $X_n^T X_n$ matrix that is near-singular, resulting in highly variable and unreliable $\hat{\beta}_{LS}$ estimates.

Further, the subset of variables to include in $X_n$ must be pre-determined by an appropriate method. As previously discussed, a primary concern with having too many covariates in a linear regression model is overfitting and modeling the noise that is specific to the dataset at hand rather than the true signal. Once included in the regression model, the $k$-th coefficient for $(k = 1, \ldots, p)$ $\hat{\beta}_{LS,k}$ summarizes the partial association of $X_k$ with the outcome, controlling for the other variables in the model. As such, coefficient size and interpretation are entirely dependent on the variables included or excluded from the specified model (Hastie et al. 2008). In settings where regressions are performed in an exploratory manner, this distinction is key.

## 4.3 LASSO: Definition

LASSO or *least absolute shrinkage and selection operator* was proposed by Tibshirani in 1996 as a method to improve model fitting within the context of least squares estimation. As mentioned in Section 3, MSE can be minimized by minimizing the estimator bias or variance. Least squares estimation is unbiased under the assumptions previously discussed, and as such, control of the MSE relies entirely on the variance of the estimator. The variance of the estimator was also previously shown to rely heavily on the matrix $X_n^T X_n$ and to be threatened by collinearity as well as the inclusion of a large number of variables $p$ in $X_n$. By allowing a small amount of bias into the estimation of the regression coefficients, penalized regression methods including LASSO aim to combat the high variability of traditional least squares regression estimates. An advantage LASSO has over similar penalized regression methods including ridge regression and elastic net is that LASSO simultaneously completes variable selection among the $p$ variables included in $X_n$. This process results in a smaller subset of coefficients.

The least squares formulation of LASSO as seen in Equation 6 is very similar to that of traditional linear least squares with a small modification. The additional term, making use of the $\ell_1$ norm, $\lambda||\beta||_1$, *penalizes* the total size of the coefficients, where $\lambda$ is the penalty applied to the total absolute value of the size of the coefficients. This is what controls the bias-variance tradeoff. As lambda increases, the coefficient estimates increase in bias by being forced to shrink towards zero. This, in turn, results in a less variable estimates $E[Y|X]$. When the penalty is large enough, continuous covariate selection occurs as coefficients drop directly to zero. The greater the penalization, the fewer predictors will be retained in the regression model and the prediction of $E[Y|X]$ wil continue to decrease in variability.

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} ||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1 \tag{9}$$

Alternatively, LASSO has been expressed as the optimization problem in Equation 10. These two formulations are equivalent. Given a particular $\lambda$ it is possible to find an $s$ that returns the same coefficient estimates, and vice versa. The parameterization provided in Equation 9 will be considered for the remainder of this paper.

$$\begin{gathered} \hat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} ||Y_n - X_n\beta||_2^2 \\ \text{s.t. } ||\beta||_1 \leq s \end{gathered} \tag{10}$$

Before exploring the properties of LASSO regression coefficients $\hat{\beta}_{LASSO}$, there are a few considerations that must be made. In order to best complete a fair penalization with LASSO regression, the variables in $X_n$ should be either standardized or measured in the same units. As $\hat{\beta}_{LASSO}$ directly corresponds to the magnitude of the variables, failure to standardize across variables will force the $\ell_1$ norm to reflect variable magnitude rather than meaningful variation. Further, as the intercept is simply a location parameter corresponding to the baseline mean of $Y_n$, its inclusion in the $\ell_1$ norm is typically unnecessary and inefficient. This can be managed in two ways: 1) through centering $Y_n$, or 2) removing $\beta_0$ from the set of penalized coefficients.

The LASSO regression method, and other penalization methods, introduce a tuning parameter $\lambda$ that is typically not known *a priori*. As with any parameter estimation, it is essential that a particular criterion is established as to what constitutes the "best" penalty size $\lambda$. While intricate theory surrounds the optimal estimation of the penalty $\lambda$ (Hastie et al. 2015), its estimation is typically completed via cross-validation, where optimality is defined as the $\lambda$ that minimizes the cross validated mean squared prediction error (CVM). Once the optimal $\lambda$ has been found, estimation of $\hat{\beta}_{LASSO}$ is found via cyclical coordinate descent, which, so long as certain "mild" conditions are met, allows for convergence to a global optimum. When considering a series of penalties, pathways coordinate descent is used instead, which assists in computational efficiency (Hastie et al. 2015).

LASSO and other shrinkage methods have a number of desirable qualities not shared by traditional least squares. For example, LASSO provides an excellent way of dealing with variables that may be correlated. The following quote from *Elements of Statistical Learning* describes the advantage of shrinkage methods in such situations as are often faced in public health data.

> When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients... this phenomenon is prevented from occurring. (Hastie et al. 2008, pp. 59)

Further, various papers have examined the asymptotic behavior of least squares regression models that make use of the variables selected by LASSO regression when the number of covariates is much larger than the number of observations ($p >> n$). These papers have found many desirable traits, such as estimator unbiasedness and convergence as $n$ goes to infinity (Liu and Yu 2013). This is an incredible improvement on least squares linear regression, which relies on the number of covariates preferably being much less than the number of observations.

## 4.4 LASSO: Concerns

While LASSO improves model fit, it decreases MSE by allowing the coefficient estimates $\hat{\beta}_{LASSO}$ to be biased. This sacrifice may concern researchers who hope to understand dose-response relationships, despite the sign of the coefficient still providing insight into a particular variable's relationship with the outcome. Estimation of confidence intervals proves challenging in that there is no closed form solution. However, work has been done exploring alternative ways and finite sample estimation, often including residual bootstrapping (Liu and Yu 2013; Hastie et al. 2015).

LASSO may handle collinearity better with respect to improved model fitting and prediction error, but collinearity between theoretically useful variables and nuisance variables is not differentiated in variable selection. As such, LASSO may choose to include the nuisance variable instead of the useful variable with which it displays high collinearity. While this may seem particularly problematic, recall the wildly and poorly determined coefficient estimates returned by least squares linear regression in this same setting. A tradeoff seems to be made between unreliable coefficient estimation and more reliable estimation but the potential swapping between a "useful" and "useless" pair of highly collinear covariates.

# 5 Application

One of the hypothesized primary deterrents to the implementation of penalized regression methods reported in Section 2 was a general lack of training in using R for statistical analyses, which is where most modern statistical packages exist (Walter and Tiermeier 2009). The primary motivation for the remainder of this paper is to present the results of an application of LASSO, including the necessary R code, in comparison to a published journal article that may have been an appropriate candidate for LASSO regression. As LASSO completes variable selection and model building simultaneously, I was interested in how similar the results of the LASSO regression, under varying conditions, would be with those found in the original analysis. Further, given the concern of overfitting with traditional least squares regression, I explore and compare the prediction error of least squares regression with that from the LASSO regression. The presence of LASSO-capable packages in STATA and R are further discussed in the R Appendix.

## 5.1 Original Analysis

The paper selected for review and comparison, "Test Score Gaps Between Private and Government Sector Students at School Entry Age in India", was published in 2014 regarding test scores and enrollment statuses of children aged 4.5 to 6 years old in India. The objective of the paper was "not to focus on why students in the private sector outperform those in the government sector but on when." (Singh 2014, pp. 32) While many previous studies have examined differences between the public and private sector with respect to children in primary school, the author observed that few had examined the association of public or private preschool attendance and future cognitive development. As such, the journal article hoped to extend the policy conversation to these prime developmental years by incorporating early childhood information into a few association-based least squares regressions. Using the YL dataset, the author found that children who attended private preschools were associated with substantially higher test scores than children who attended public preschools. It was then demonstrated that while much of the performance gap could be removed by controlling for parental background and particular child characteristics, the gap remained significant, warranting further investigation.

Correspondingly, the original paper considered two least squares regression models. First, an exploratory model of regressing student test scores $Y_i$ on enrollment status $enrol_i$, described by Equation 11, controlling for site effects $site_i$. The inclusion of $site_i$ as a control was critical for the differences between sites, as described in the original paper:

> Site fixed effects, included here by including a vector of site dummies, allow for removing any levels differences between different sites (mandals). This is particularly important in this case since the take-up of different institution types varies much across the sites in the sample. In effect, these regressions compare children who are in different types of education but living in the same cluster (Singh 2014, pp. 46).

$$Y_i = \alpha + \beta_1 \mathbf{enrol}_i + \beta_2 \mathbf{site}_i + \epsilon_i \tag{11}$$

The second model added to the first a vector $X_i$ of potentially confounding covariates. There was no specification as to how this subset of potential confounding covariates were selected, except that they were intended to capture the socioeconomic background of the children and their families. This model is defined in Equation 12.

$$Y_i = \alpha + \beta_1 \mathbf{enrol}_i + \beta_2 \mathbf{site}_i + \beta_3 X_i + \epsilon_i \tag{12}$$

Enrollment status, $enrol_i$, was further defined and tested in two different ways. First, each of the models were run only considering current enrollment status as the exposure of interest. Then, enrollment status was re-formatted to include the student's enrollment history up to and including their current enrollment status. Each of these possible combinations are outlined in the codebook in Table **??**.

## 5.2 Candidacy for LASSO

The objective of the original paper has several qualities that may be difficult to efficiently manage through traditional regression methods. First, this is a high dimensional problem. While there is a large sample size of 1,941 children with complete case information for CDA scores and 1,829 children with complete PPVT information, there are considerably more variables measured in the YL dataset, and at a multitude of hierarchical levels. In this setting (children within school systems within sites), even completing variable selection based on a rigorous causal diagram could become very intricate, with the potential for overadjustment. Additionally, the author was consistenty clear as to the exploratory rather than confirmatory nature of the analysis in the original paper. As such, one may be interested in a multitude of potential covariates rather than a narrow testing subset. Finally, the potential for highly collinear covariates, especially considering the clustering, should be of great concern for the stability of traditional regression estimates.

Using LASSO, I have two primary goals. First, I want to simplify the model building process. As this is an exploratory investigation, I would like to consider multiple variables while avoiding *ad hoc* inclusion and exclusion of the variables in the model. Of course, in order to stay true to the models presented in the original analysis, I am unable to include and explore the findings of LASSO in a large and unexplored subset of covariates, though I do attempt a version of exploratory analysis in the first model considered, which has known high collinearity. Second, given the concerns about the performance of a least squares estimator in this setting, I hope to produce a less variable estimator of student performance. To complete these goals, I investigate two main questions of interest. First, I build an exploratory model that includes the data-driven strongest predictors of cognitive measures in the YL cohort of children from India. Then, I replicate the Model 12 efforts by forcing the exposure of interest to remain in the model while allowing LASSO to select which additional covariates to include from those predetermined by the original author. Model 11 is not explicitly replicated, but can be inferred from the variable trace plots in Figures 3 and 4. In replicating Model 12 with

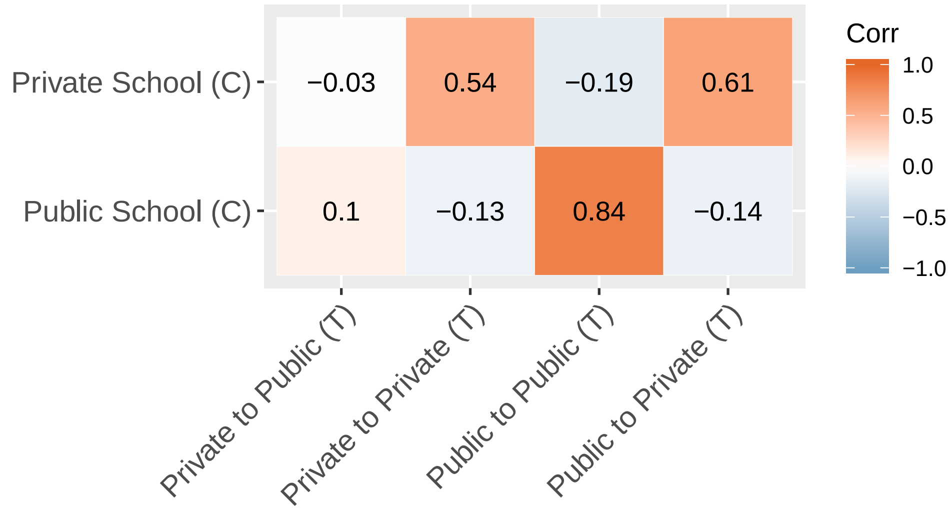



Figure 1: Map of the correlation between a subset of the variables describing the two measures of enrollment: 1) current enrollment status and 2) enrollment trajectory. It appears that students who are currently in public primary school typically attended a public preschool earlier in life and rarely a private preschool. Conversely, students enrolled in private primary school appear to have a variety of backgrounds, with both public and private preschools displaying a strong correlation.

LASSO, I am prioritizing the inclusion of variables that have a large effect on untangling the signal from the noise in the data.

## 5.3 LASSO Model

### 5.3.1 Exploratory Model

My first exploratory question was of all of the variables of interest specified in the original analysis, what are the strongest predictors of cognitive measures for the YL cohort in India? To this end, I ran a LASSO regression on all variables used in the original models, indescriminant of any presupposed exposure relationship with the outcome of interest. As such, the potential pitfall of this model is that LASSO may not select enrollment, in either formulation, as a significant predictor of the outcome (standardized CDA and PPVT scores). Therefore, any coefficient comparison with the original analysis would be misleading as the models may include different variables.

Including both current enrollment and enrollment trajectory in the model induces a high level of collinearity among the variables included in the model. As can be seen in Figure 1, students who are currently attending a private primary school seem to have trajectories that include public preschools (correlation of 0.61) as well as private preschools (0.54). Conversely, students currently enrolled in public schools very rarely had a private preschool education (0.1) versus a public preschool education (0.84) or no preschool education at all (0.36, not shown). These high correlations between current enrollment status and trajectory may be important for further consideration if the goal is to identify the true support of the LASSO regression.

Moving forward with the analysis, the largest penalty that produces a CVM within one standard deviation of the minimum CVM is applied. Under this penalty, the predictors and coefficients LASSO returned as important when considering CDA scores or PPVT scores as the outcome are presented in Table 2. A separate column for coefficient signs has been included for ease of comparison. The results for modeling PPVT and CDA scores are included in the same table for clarity of overlap and consistency of coefficient sign. The code used in this and later analyses can be found in the R Appendix.

In all of the models, the baseline educational status was considered to be no school or unreported school status (CDA: n = 130, PPVT: n = 119). Notice, current enrollment in a public preschool or public primary school, as well as the trajectories from private preschool to private primary school and public preschool to private primary school were not returned as nonzero coefficients for either outcome and, therefore, are not included in this table. Since current enrollment in a public preschool consistently has one of the smallest and least significant coefficient estimates across all of the methods and models considered, it is relatively unsurprising that the coefficient under this strict penalty is dropped to zero. However, dropping the coefficient of the private preschool to private primary school trajectory warrants further examination. First, this may be a result of the high correlation between the trajectory and current enrollment status variables. Additionally, the coefficient trace plots produced by LASSO can be invaluable to this end. The trace plots for this model in Figure 2, demonstrate that the trajectories of private preschool to private school trajectory (traj7) and public preschool to private primary school (traj9) under the optimal CVM penalty as well as the least squares regression output in (Tables 5 and 6) have some of the largest coefficient estimates of any of the covariates included in the models. The restrictive nature of the penalty chosen in the model, while beneficial for inducing sparsity in coefficient estimates, may be obscuring an important relationship in this case.

Looking at the results more generally, there is a positive association between test scores and current private school enrollment or any history of private school education. Further, when parents are well educated, their children appear to score higher on tests, confirming results from existing literature (Davis-Kean 2005). Examination of the coefficient trace plots in Figure 2 presents a clear majority of these coefficients maintaining a consistent positive or negative association with the test scores, regardless of penalty size. Without further analysis, this exploratory model seems to present similar results as those found by the original paper. As a note, site level effects have not been displayed for any of the models as they were not the primary interest of the original analysis.

### 5.3.2 Comparative Models

While the exploratory analysis provides insight into the associations between the variables of interest and the outcome, it is not comparable to the original analysis. As such, a LASSO model including strictly the covariates specified in Equation 12 was fit, forcing the regression to keep enrollment status, in its proper formulation, in the model. A first LASSO regression considered the exposure to be enrollment status as represented by current enrollment. Next, a model was fit using the observed trajectory of the students' enrollments as the exposure

| | CDA | | PPVT | |
|---|---|---|---|---|
| Covariate | Signs | Coefficient | Signs | Coefficient |
| Public Primary School (Current) | + | 0.0580 | + | 0.0515 |
| Private Primary School (Current) | + | 0.3946 | + | 0.4361 |
| Private Preschool Only$^t$ | + | 0.2305 | + | 0.2338 |
| Public Preschool Only$^t$ | - | -0.1861 | | |
| Private Primary School Only$^t$ | + | 0.1264 | + | 0.0724 |
| Private Preschool to Public Primary School$^t$ | + | 0.1905 | | |
| Public Preschool to Public Primary School$^t$ | + | 0.0051 | + | 0.1370 |
| Household Size | - | -0.0139 | - | -0.0115 |
| Father's Education | + | 0.0216 | + | 0.0201 |
| Mother's Education | + | 0.0514 | + | 0.03978 |
| Child's Age (Months) | - | -0.0033 | - | -0.0030 |
| Female-headed Household | - | -0.2215 | - | -0.1977 |

$^t$ Trajectory variables

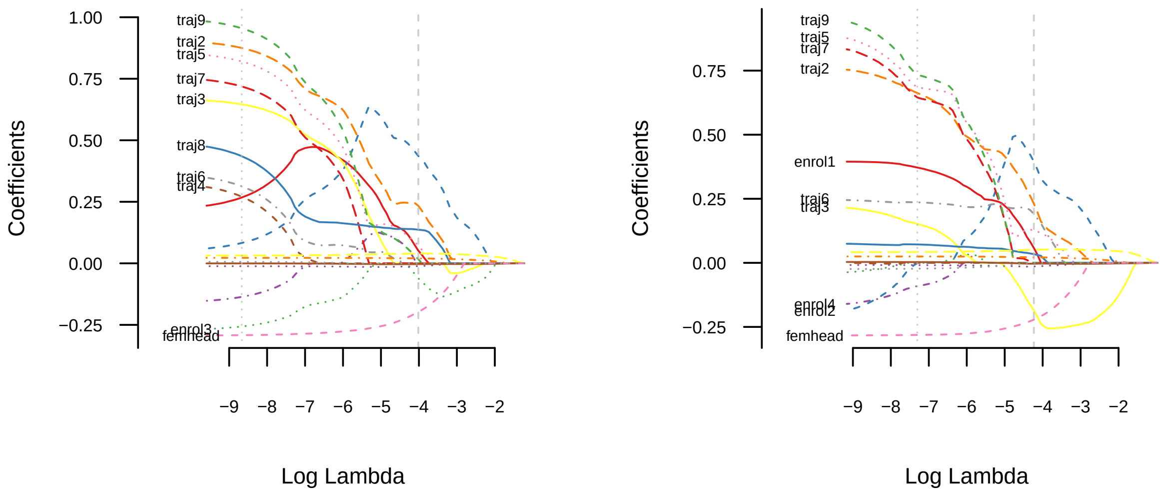
Table 2: The coefficient estimates and signs of the nonzero coefficients returned by the exploratory LASSO model when CDA and PPVT are used as the respective outcomes of interest. Both models consisted of CDA or PPVT as a function of the two forms of enrollment, household size, father's and mother's education status, the child's age in months, whether the child was the firstborn, and whether the household was headed by a female. While the models for CDA and PPVT may not have included the exact same nonzero coefficients, there is perfect correspondence regarding the positive or negative contribution of each of the shared coefficients with the outcome.

of interest.

Tables 3 and 4 compare the coefficient estimates reported in the original analysis to the coefficient estimates provided by LASSO for the models considering current enrollment status as the primary exposure of interest on cognitive scores. Results from the LASSO regression when the optimal penalty (that which minimizes the CVM) and the more strict penalty (that which produces results within one standard error of the minimum CVM) are provided. A consistent relationship is observed across all models for a majority of the variables. Larger household sizes and female-headed households appear to be negatively associated with PPVT and CDA test scores. Conversely, enrollment at a private preschool or private primary school continues to have the largest estimated positive association with high test scores.

The coefficient trace plots for these models are provided in Figures 3 and 4. While forcing the exposure to remain in the model creates a strange shape for the exposure variables, the coefficient estimates again remain mostly stable with respect to their positive or negative association with cognitive scores regardless of penalty size.

This “near-ness” of the two methods (traditional least squares and LASSO) is unsurprising given the stability of the coefficient traces in Figures 3 and 4 across penalty sizes. As a rather small set of covariates are being considered in these models, large differences in coefficient estimates are unlikely. Even those covariates that are not forced to remain in the model are relatively stable in their size and direction of association with the outcome. In




CDA: Exploratory LASSO

PPVT: Exploratory LASSO

Figure 2: Trace plot for the coefficient estimates across a range of penalties from least to most restrictive. Both models consisted of CDA or PPVT as a function of the two forms of enrollment, household size, father's and mother's education status, the child's age in months, whether the child was the firstborn, and whether the household was headed by a female. The dotted vertical line represents the penalty at which the CVM was minimized. The dashed vertical line represents the most restrictive penalty that returns a CVM within one standard error of the minimum.

| **Coefficient** | **LS** | **LASSO (Optimal)** | **LASSO (1SE)** |
|---|---|---|---|
| Public school | 0.46*** | 0.447 | 0.371 |
| Private school | 0.72*** | 0.706 | 0.752 |
| Public preschool | 0.28*** | 0.168 | 0.07 |
| Private preschool | 0.53*** | 0.488 | 0.447 |
| Household Size | -0.0031 | -0.011 | -0.001 |
| Mother's Education Level | 0.034*** | 0.033 | 0.034 |
| Father's Education Level | 0.023*** | 0.024 | 0.016 |
| Age in Months | 0.035*** | 0.002 | -0.007 |
| Female-headed Household | -0.29*** | -0.3 | -0.107 |
| First-born Child | 0.024 | 0.01 | 0 |
| Constant | -2.98*** | 0 | 0 |
| Observations | 1941 | 1940 | 1940 |

Table 3: Coefficient estimates for the least squares and LASSO models regarding current enrollment status' association with CDA scores and controlling for the covariates listed in the table. Original least squares (LS) results source: *Test Scores Gaps Between Private and Government Sector Students at School Entry Age in Inda*, Singh (2014), pp. 41.

| **Coefficient** | **LS** | **LASSO (Optimal)** | **LASSO (1SE)** |
|---|---|---|---|
| Public school | 0.38*** | 0.364 | 0.339 |
| Private school | 0.57*** | 0.594 | 0.72 |
| Public preschool | 0.16 | 0.08 | 0.009 |
| Private preschool | 0.43*** | 0.428 | 0.434 |
| Household Size | -0.0046 | -0.01 | -0.013 |
| Mother's Education Level | 0.043*** | 0.044 | 0.051 |
| Father's Education Level | 0.025*** | 0.026 | 0.02 |
| Age in Months | 0.031*** | 0.004 | -0.007 |
| Female-headed Household | -0.28** | -0.302 | -0.209 |
| First-born Child | -0.0098 | -0.016 | 0 |
| Constant | -2.69*** | 0 | 0 |
| Observations | 1829 | 1828 | 1828 |

Table 4: Coefficient estimates for the least squares and LASSO models regarding current enrollment status' association with effects on PPVT scores and controlling for the covariates listed in the table. Original least squares (LS) results source: *Test Scores Gaps Between Private and Government Sector Students at School Entry Age in Inda*, Singh (2014), pp. 41.

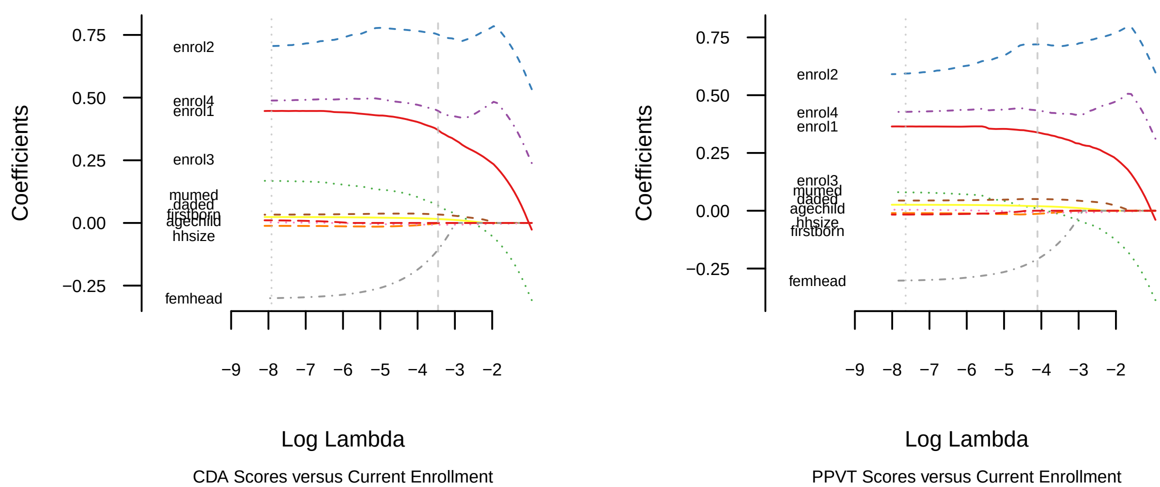



Figure 3: Trace plots for the coefficient estimates across a range of penalties from least to most restrictive. Both models consisted of CDA or PPVT as a function of current enrollment status, household size, father's and mother's education status, the child's age in months, whether the child was the firstborn, and whether the household was headed by a female. The dotted vertical line represents the penalty at which the CVM was minimized. The dashed vertical line represents the most restrictive penalty that returns a CVM within one standard error of the minimum. Both models considered the exposure of interest to be current enrollment status.

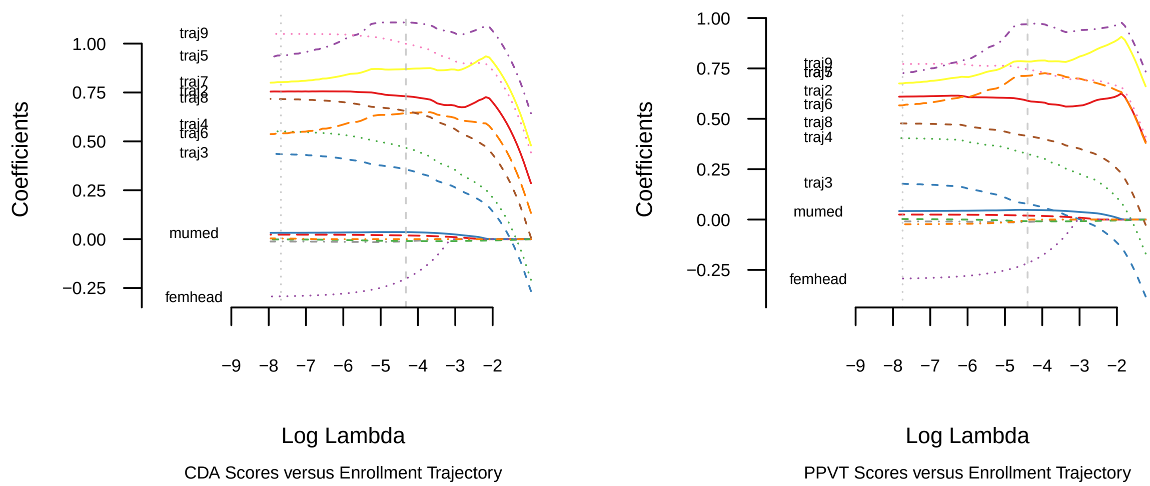



Figure 4: Trace plot for the coefficient estimates across a range of penalties from least to most restrictive. Both models consisted of CDA or PPVT as a function of enrollment trajectory, household size, father's and mother's education status, the child's age in months, whether the child was the firstborn, and whether the household was headed by a female. The dotted vertical line represents the penalty at which the CVM was minimized. The dashed vertical line represents the most restrictive penalty that returns a CVM within one standard error of the minimum. Both models considered the exposure of interest to be the observed enrollment trajectories of the students.

| **Coefficient** | **LS** | **LASSO (Optimal)** | **LASSO (1SE)** |
|---|---|---|---|
| Private Preschool Only | 0.85*** | 0.755 | 0.731 |
| Public Preschool Only | 0.61*** | 0.435 | 0.356 |
| Public School Only | 0.61*** | 0.551 | 0.467 |
| Private School Only | 1.06*** | 0.94 | 1.108 |
| Private Preschool to Public School | 0.65** | 0.541 | 0.644 |
| Private Preschool to Private School | 0.85*** | 0.803 | 0.87 |
| Public Preschool to Public School | 0.78*** | 0.716 | 0.655 |
| Public Preschool to Private School | 1.12*** | 1.049 | 0.999 |
| Household Size | -0.0024 | -0.012 | -0.011 |
| Mother's Education Level | 0.034*** | 0.032 | 0.036 |
| Father's Education Level | 0.023*** | 0.023 | 0.019 |
| Age in Months | 0.035*** | -0.001 | -0.01 |
| Female-headed Household | -0.29*** | -0.292 | -0.202 |
| First-born Child | 0.017 | 0.003 | 0 |
| Constant | -3.27*** | 0 | 0 |
| Observations | 1941 | 1940 | 1940 |

Table 5: Coefficient estimates for the least squares and LASSO models regarding CDA as the outcome and enrollment trajectory as the exposure of interest and controlling for the covariates listed in the table. Original least squares (LS) sesults source: *Test Scores Gaps Between Private and Government Sector Students at School Entry Age in Inda*, Singh (2014), pp. 42.

this example the trade-off in estimate bias appears nearly negligible, even when considering stricter penalty sizes. The same results hold when considering trajectory of enrollment as the primary exposure of interest. Again, the LASSO regression produces similar results to those from the least squares regression, as can be seen in Tables 5 and 6.

Because the coefficient estimates are rather similar and stable, the changes in estimate should reflect a decrease in the variability of the LASSO estimator. In essence, because LASSO trades bias for variance, it produces an estimator that is more stable and less influenced by outliers or extreme values. Figure 5 examines this relationship by plotting both the predicted mean squared errors for the least squares model and those for the LASSO models considered. A subset of 80% of the data was used to fit each model while the remaining 20% was used to determine the predicted MSE that is shown in the plots. Figure 5 demonstrates that every model of LASSO outperforms the least squares model. This, as mentioned, is because LASSO takes advantage of the bias-variance tradeoff to minimize the predicted MSE, producing less variable estimates of the outcome and as such, performing better when faced with new data.

Further, in Figure 5 LASSO* refers to the exploratory models that contained both the current enrollment status as well as the trajectory of enrollment for each student. As such, this model is guaranteed to contain highly collinear variables. Unlike traditional regression methods, where this would result in highly variable estimates, these models are still among

| **Coefficient** | **LS** | **LASSO (Optimal)** | **LASSO (1SE)** |
|---|---|---|---|
| Private Preschool Only | 0.65*** | 0.61 | 0.589 |
| Public Preschool Only | 0.30* | 0.178 | 0.077 |
| Public School Only | 0.47*** | 0.404 | 0.325 |
| Private School Only | 0.79*** | 0.726 | 0.969 |
| Private Preschool to Public School | 0.62** | 0.567 | 0.714 |
| Private Preschool to Private School | 0.69**** | 0.676 | 0.784 |
| Public Preschool to Public School | 0.53*** | 0.477 | 0.414 |
| Public Preschool to Private School | 0.80*** | 0.772 | 0.745 |
| Household Size | -0.0030 | -0.009 | -0.012 |
| Mother's Education Level | 0.043*** | 0.042 | 0.048 |
| Father's Education Level | 0.025*** | 0.025 | 0.02 |
| Age in Months | 0.030*** | 0.003 | -0.008 |
| Female-headed Household | -0.28** | -0.293 | -0.216 |
| First-born Child | -0.019 | -0.023 | -0.002 |
| Constant | -2.76*** | 0 | 0 |
| Observations | 1829 | 1828 | 1828 |

Table 6: Coefficient estimates for the least squares and LASSO models regarding PPVT as the outcome and enrollment trajectory as the exposure of interest and controlling for the covariates listed in the table. Original least squares (LS) results source: *Test Scores Gaps Between Private and Government Sector Students at School Entry Age in Inda*, Singh (2014), pp. 42.

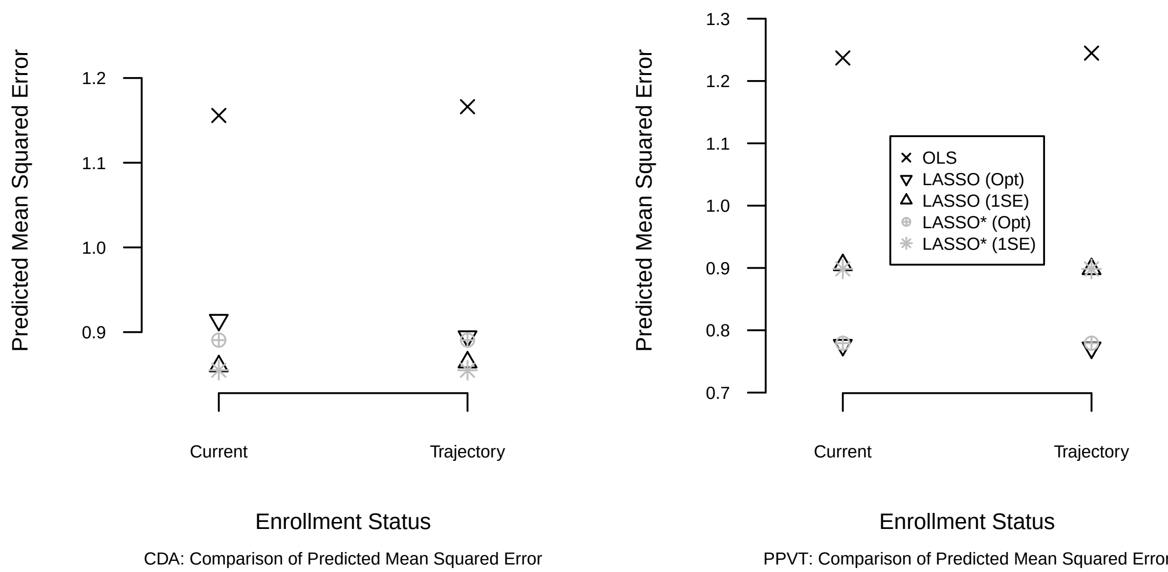



Figure 5: The predicted mean squared error for each of the models considered. "Opt" denotes the results from the LASSO penalization that minimized the CVM on the training set. "1SE" denotes the results from the strictest LASSO penalization that returned a CVM on the training set within one standard error of the minimum.

the best performing. This speaks directly to the ease with which LASSO handles collinear data.

## 5.4 Summary

To briefly summarize, in this setting, the LASSO regression primarily functioned as a manner of decreasing the variance of the estimator. As the subset of covariates had already been chosen to be rather selective, variable selection by way of penalization was rather unnecessary. This is not to say that covariates were not dropped. In all of the models the coefficient corresponding to whether the student was the first-born child was dropped to zero in the strictest penalization. Decreased estimator variability, in and of itself, is desirable for a variety of reasons including the avoidance of overfitting. As such, LASSO in this application setting was still a valuable exercise as it successfully produced estimators that performed better when faced with new data.

Through a content focused lens, LASSO points to a rather stable association between enrollment in private education and higher cognitive outcomes as measured by PPVT and CDA scores. Beginning with the uni-directional correlations between current enrollment status and student enrollment trajectories and continuing through each of the models, the theme of a desirable private school education and its association with higher scores emerges.

Though both analyses have been completed in a strictly exploratory way, the results certainly warrant further investigation.

# 6 Discussion

The application in this paper demonstrated a simple three step process to complete LASSO estimation. First, it is necessary to determine what will be considered by the model. In this application, and in other non-experimental studies like it, this selection is one point at which existing literature, prior knowledge and statistical tools such as DAGs can be incredibly informative. A concern with statistical methods designed for high dimensional data is that the method is simply a "black-box". By being explicit in the the theories and existing understanding which is used to inform which variables or types of variables are considered for selection in the model, this "black-box" concern can begin to be satisfied. This is a highly recommended practice even for traditional analyses, particularly in non-experimental studies such as those generated from large pre-existing databases (Sauer, Brookhart and Roy 2013). Second, cross-validation can be used to determine an optimal penalty. Cross-validation ascertains that the data used to build the model is different than the data used to test the model's performance. This prevents the application of a penalty that is strictly optimal to the current data itself, i.e. avoiding overfitting. Third, LASSO regression is completed and coefficient estimates can be returned using the optimal penalty or another useful penalty such as those that induce greater sparsity, such as the strictest penalty which returns a CVM within one standard error of the minimum. As seen in Figure 5, least squares regression tends to fit the data well, but tends to produce highly variable estimates when faced with new data. The shrinkage LASSO enforces, while creating bias in the coefficient estimates themselves, produces less variable predictions when faced with a new dataset, therefore minimizing the overall MSE. In three lines of code, it is possible to produce a model that, contrary to least squares, 1) is more rigorous to collinearity, 2) can select a subset of covariates even when there are more covariates than observations ($p >> n$), and 3) produces less variable estimates.

A next step would be to consider the blossoming field of high dimensional and post model selection inference and how this pertains to LASSO coefficient estimates. In the most traditional regression analyses, the subset of covariates considered by a model are specified *a priori*. Data is then collected pertaining to these variables and a model is fit. In data-adaptive methods, including step-wise, the covariates included in the model are chosen based on the data. As such, the inference regarding their coefficient estimates should reflect the fact that they have already been "tested" in a way that determined whether or not the variable remained in the model set. It is this difference, a prespecified set of covariates versus an adaptive set of covariates, that makes traditional applications of inference incorrect and misleading. As such, a number of methods have been suggested for use with LASSO regression, though a consensus has not been met as to which method may be most favorable. *Statistical Learning with Sparse Data* (Hastie et al. 2015) describes several such methods ranging from Bayesian models to bootstrapping. As of the writing of this paper, a method for obtaining p-values within the `glmnet` package was not yet implemented.

In conclusion, variable selection is an important part of completing reliable estimation. As such, journals should require greater transparency in order to advance the field as a whole. Many frequently used methods are highly contested for well-documented reasons. Methods that were once optimal because of properties such as low computational expense or the sense of stability that comes from completing an ordered set of stepwise tests are losing relevance to newer methods that make better use of modern computing power, respond better to the demands of high dimensional datasets and help remove the researcher bias of cherry-picking model results that fit the research agenda. LASSO is simply one of many methods that aims to intelligently combine methods, such as least squares, that are well known and understood with the need for high dimensional variable selection. In order to perform lasting research and avoid the growing concerns of null findings (Ioannadis 2005), public health research should continue to pursue greater transparency and the implementation of improved methods.

# References

Angrist, J.D. and Pischke, J. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princenton University Press.

Christensen, R.A. (1987), *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer-Verlag New York Inc., New York, NY.

Davis-Kean, P.E. (2005), "The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment," *Journal of Family Psychology*, (Vol. 19, No. 2), 294-304.

Dziak, J., Li, R. and Collins, L. (2005) "Critical Review and Comparison of Variable Selection Procedures for Linear Regression," Technical Report, Methodology Center, Penn State University.

Greenland, S. (2008), "Invited Commentary: Variable Selection versus Shrinkage in the Control of Multiple Confounders," *American Journal of Epidemiology*, (Vol. 167, No. 5), 523-531.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY.

Ioannidis, J. P. A., (2005), "Why Most Published Research Findings are False," *Plos Medicine*, (Vol. 2, No. 8).

Lee, P. H. (2014), "Is a Cutoff of 10% Appropriate for the Change-in-Estimate Criterion of Confounder Identification?" *Journal of Epidemiology*, (Vol.24, No.2), 161–167.

Liu, H., Yu., B., et. al. (2013), "Asymptotic Properties of LASSO+mls and LASSO+ridge in Sparse High-Dimensional Linear Regression," *Electronic Jounal of Statistics*, (Vol. 7), 3124-3169.

Logo, D. L. and Drazen, J. M. (2016), "Data Sharing," *New England Journal of Medicine*, (Vol 374, No. 3), 276-277.

Robins, J. M. (1987), "A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods," *Journal of Chronic Disease* (Vol. 40, Supplement), 2:139s-161s.

Robins, J. M. and Greenland, S. (1986). "The Role of Model Selection in Causal Inference from Nonexperimental Data," *American Journal of Epidemiology.* (Vol. 123, No. 3), 392-402.

Rosenbaum, P. R., and Rubin, D. B. (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, (Vol. 70, No. 1), 41-55.

Sauer, B., Brookhart, M.A. and Roy, J.A., et al. (2013), "Covariate Selection," *Developing a*

*Protocol for Observation Comparative Effectiveness Research: A User's Guide*, ed. Velentgas P., Dreyer N.A., Nourja P., et al., Rockville, MD.

Schisterman E. F., Cole S. R., and Platt, R. W. (2009), "Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies," *Epidemiology* (Vol. 20, No. 4), Cambridge, MA, 488-495.

Singh, A. (2014) "Test Score Gaps between Private and Government Sector Students at School Entry Age in India," Oxford Review of Education, (Vol. 40, No. 1), 30-49.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J. and Schneeweiss, S. (2006) "A Review of the Application of Propensity Score Methods Yielded Increasing Use, Advantages in Specific Settings, but not Substantially Different Estimates Compared with Conventional Multivariable Methods," *Journal of Clinical Epidemiology*, (Vol. 59, No. 5), 437-47.

Walter, S. and Tiemeier, H. (2009), "Variable Selection: Current Practice in Epidemiological Studies," *European Journal of Epidemiology* (Vol. 24, No. 12), 733-736.

Wasserman, L. (2010), *All of Statistics: A Concise Course in Statistical Inference*, Springer Publishing Company Incorporated.

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statment on P-Values: Context, Process, and Purpose," *The American Statistician*, (Volumne 70, No. 2), 129-133.