Model Selection and Shrinkage: A Project in Linear Modeling

David DeVine

1 Introduction

Linear Modeling, in a broad sense, involves using information that we have to predict outcomes in the future. As one can gather, this is concept is widely applicable to disciplines across the board.

Big data is an up and coming catch phrase in the media these days. With the rise of technology and the Internet, we have more data at our fingertips than ever before. The question then becomes: How can we most effectively use it to our advantage? How can we filter it so that it is managable and still useful to us? How can we profit from or make use of its patterns? Data is cheap and easy to find. Effective tools for analysis and individuals who can use these tools, on the other hand, are neither cheap nor easy to find.

This is where linear modeling comes in. Statisticians are working to develop algorithms that will efficiently select, compile, and sort through these data sets to most efficiently flesh out pertinent information. Using model selection, statisticians are able to parse through data sets, emphasizing the most pertinent information and diminishing or ignoring information which is repetitive or unrelated. In this way, our data sets are both interpretable and more computationally feasible so that prediction can move with the click of a button.

As will be discussed in this paper, the most popular linear modeling technique is the Ordinary Least Squares technique, which minimizes the sum of the squared errors of the model. This technique, however, is unreliable when there are extreme outliers or when there is collinearity between variables in the model. Statisticians have been developing new modeling techniques to remedy these issues. This paper will explore the machinery behind these techniques, study their asymptotic behaviors, and run simulations to analyze their performance in various modeling scenarios.

2 Linear Models

2.1 Some Notation

As will be discussed, there exist many different modeling techniques at the fingertips of statisticians, each with a different approach to the problem at hand and each with their own pros and cons. The goal of the statistican is to use the one which is best. An illustrative example will help us see what this could mean.

Suppose we want to model systolic blood pressure against age. In other words, this model will predict an individual's blood pressure given their age. We collect data from existing patients and plot it, as in Figure 5. The line through the data shows one example of a predictive linear fit of the data. Input an age and the curve will output an expected blood pressure.

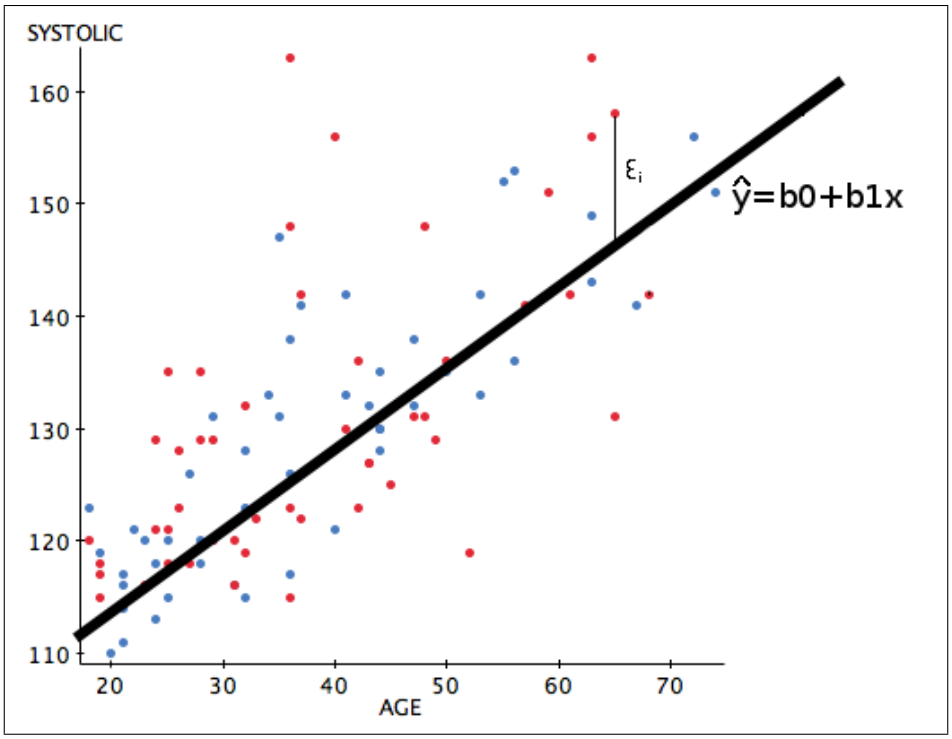

Figure 1: Blood Pressure vs. Age

One obvious criterion for the performance of a linear model is to sum up the squares of the residuals, shown in the figure to be the distance between the predictive curve and each individual data point. We will discuss this in more detail later.

In the above example we collected data on one *predictor* (age) and one *outcome* (blood pressure). In general , we have p predictors and list their observations in the *obervational matrix* as below.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & . & . & . & x_{p1} \\ 1 & x_{12} & x_{22} & . & . & . & x_{p2} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & . & . & . & x_{pn} \end{bmatrix}$$

The matrix X is $n \times P$ where $P = p + 1$. The first column of ones accounts for an intecept term β_0. We assume that the outcomes y_j are distributed as

$$y_j = \beta_0 + \beta_1 x_{1j} + ... + \beta_p x_{pj} + \epsilon_j = X_j^T \beta + \epsilon_j \tag{1}$$

$1 \leq j \leq n$. The vector β is the $P \times 1$ vector $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ and the ϵ terms are normally distributed with a mean of 0 and a standard deviation of σ. We write: $\epsilon \sim \mathcal{N}(0, \sigma)$.

All modeling techniques discussed attempt to estimate the value of β through some estimator $\hat{\beta}$. The hat notation indicates that the quantity is an estimate of β. Given a set of n observations, we apply a modeling technique which generates an estimate $\hat{\beta}$ of the true β. Because $\hat{\beta}$ depends on a set of observations, its value will change as the set of observations changes. In this sense, the estimator $\hat{\beta}$ is a random variable. Often times we are concerned with the *bias* of such an estimator which captures the difference between the expected value of an estimator and the true value of the parameter which it estimates.

In general, we say that an estimator, $\hat{X}$ is *unbiased* iff $E(\hat{X}) = X$.

This leads right in to the *variance* of an estimator, defined as the expected value of the squared sampling deviations, that is,

$$var(\hat{X}) = E[(\hat{X} - E(\hat{X}))^2] = \sigma^2(\hat{X})$$

This tells us the average squared distance an estimate $\hat{X}$ is away from its expected value, capturing a measure of spread of the values of $\hat{X}$.

We also examine the covariance matrix of the predictors. The covariance between two variables X_1 and X_2 is measured by

$$\sigma(X_1, X_2) = E\{[X_1 - E(X_1)][X_2 - E(X_2)]^T\}.$$

Covariance measures the percentage of variation in one variable that is explained by variation in the other. Thus we define the covariance matrix of the set of variables by

$$\sigma^{\mathbf{2}}(\mathbf{X}) = \mathbf{E}\left(\begin{bmatrix} X_1 - E(X_1) \\ X_2 - E(X_2) \\ \cdot \\ \cdot \\ \cdot \\ X_p - E(X_p) \end{bmatrix} \begin{bmatrix} X_1 - E(X_1) & X_2 - E(X_2) & ... & X_p - E(X_p) \end{bmatrix}\right)$$

$$\sigma^{\mathbf{2}}(\mathbf{X}) = \begin{bmatrix} \sigma^2(X_1) & \sigma(X_1, X_2) & ... & \sigma(X_1, X_p) \\ \sigma(X_2, X_1) & \sigma^2(X_2) & ... & \sigma(X_2, X_p) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \sigma(X_p, X_1) & \sigma(X_p, X_2) & ... & \sigma^2(X_p) \end{bmatrix}.$$

Noting $\sigma(X_1, X_2) = \sigma(X_2, X_1)$ we see that covariance matrices are symmetric. These matricies are helpful at identifying highly correlated variables. Note that the *correlation* of X_1 and X_2 is defined as $corr(X_1, X_2) = \sigma(X_1, X_2)/\sqrt{\sigma^2(X_1)\sigma^2(X_2)}$. Correlation matricies are defined similarily and serve to standardize covariance matricies by making the diagonal entries equal to 1.

The correlation matrix captures pairwise linear correlations between the predictor variables. For a more intuitve sense of this, we can imagine that in predicting a level of job satisfaction, *years of*

experience and *annual income* would be highly correlated predictors: when one changes, the other tends to change as well. This idea of correlation between predictors will play a large role in the performance of modeling techniques.

2.2 Ordinary Least Squares

The Ordinary Least Squares (OLS) technique is the oldest and most popular method by which we measure other modeling techniques. Many of the techniques discussed in the paper involve some type of modification to the least squares approach. We will study it now to establish its machinery and its performance so that we have a basis for comparsion.

The OLS approach attempts to minimize the squares of the error terms $\epsilon_i = y_i - X_i^T\beta$ as such:

$$\hat{\beta}^{OLS} = \min_{\beta} \sum_{i=1}^{n} (\epsilon_i)^2$$

where y_i is the predicted outcome for the ith set of predictor observations x_i. Minimizing this quantity over all possible values of β will produce an optimal $P \times 1$ vector $\hat{\beta}^{OLS}$ which satisfies

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

and the OLS model then becomes

$$\hat{y} = X\hat{\beta}^{OLS}$$

where again, a hat is placed over the y to indicate that it is only an estimate of our true observations y. We can see that our estimates are dimensionally correct by noting that the matrix X is $n \times P$ and the vector $\hat{y}$ is $n \times 1$. Each element of $\hat{\beta}$ represents the estimated coefficient of a corresponding element of the true β. To foreshadow a future discussion, we note that if the matrix $X^T X$ is not invertible then problems arise in our $\hat{\beta}^{OLS}$ estimates.

Theorem 1. *Suppose y_j is distributed as in (1). Then, $\hat{\beta}^{OLS}$ is an unbiased estimator of β.*

Proof. By definition, $\hat{\beta}^{OLS} = (X^T X)^{-1} X^T (X\beta + \epsilon)$. Thus,

$$\begin{aligned} E[\hat{\beta}^{OLS}] &= E[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\ &= E[\beta] + E[(X^T X)^{-1} X^T \epsilon] \\ &= \beta + (X^T X)^{-1} X^T E[\epsilon] \\ &= \beta + 0 = \beta. \end{aligned} \tag{2}$$

By noting that E is a linear operator with $E[\epsilon] = 0$ and that x_i is not a random variable by assumption. □

This is reassuring. It tells us that, on average, the $\hat{\beta}$ coefficients will correctly estimate the β coefficents in the true model. On the other hand, the variance of $\hat{\beta}$ can be large. Since variance captures the spread of an estimate, we can say that high variance in the $\hat{\beta}$ coefficents results

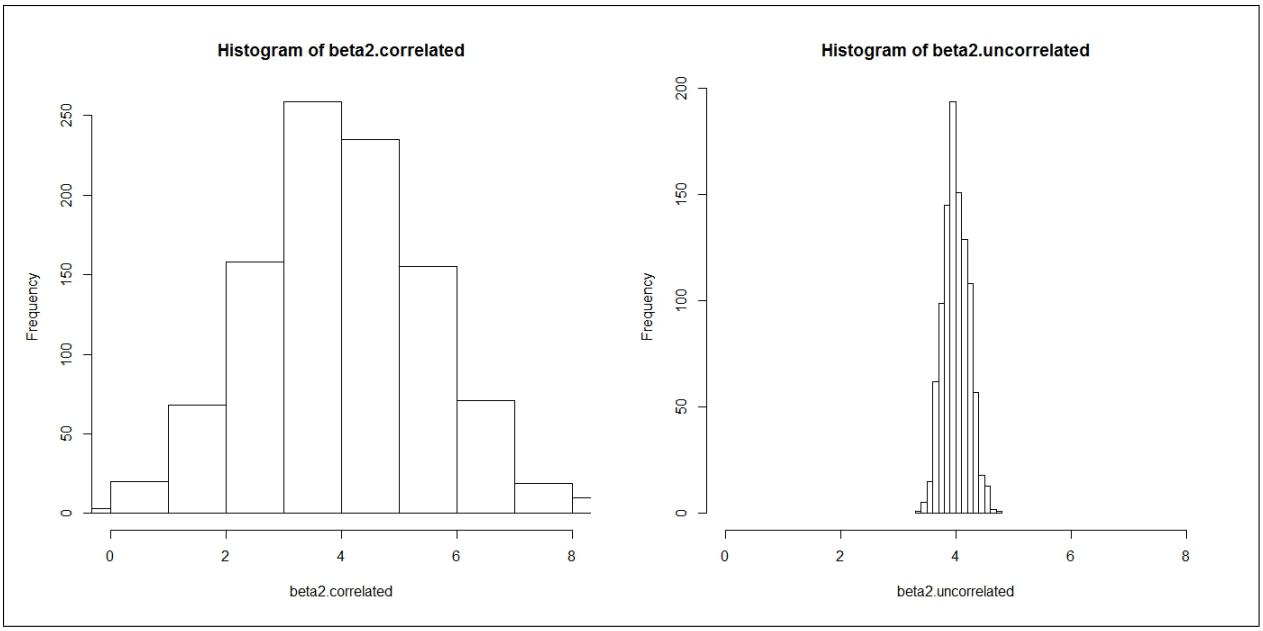

Figure 2: Sampling Variations in $\hat{\beta}$ against correlation between predictors

in imprecise estimates of β. So although OLS produces unbiased esimates it can also have high variance in the estimates, causing them to be unreliable.

When does the OLS estimate have high variance? As we will see, it will be when the predictors are highly correlated. Earlier we noted that the $\hat{\beta}^{OLS}$ are highly sensitive to the invertibility of $X^T X$. We review the following theorem from Lay [5] to find where the OLS technique is unreliable.

Theorem 2. *The matrix $A^T A$ is invertible iff the columns of A are linearly independent. [5]*

If we have highly correlated predictors then there exists some dependancy between the colums of X. In this sense, there is a spectrum of the invertibility of a matrix. This causes interesting dynamics in the $\hat{\beta}$ values. To illustrate this we ran some simulations in **R** to generate histograms of estimator coefficients in two models. Both models had two predictors (p=2). The first model had highly correlated predictors $corr(X_1, X_2) = .99$, while the second had lower correlation between predictors $corr(X_1, X_2) = .2$. The histograms in Figure 2 above reveal a widely spread distribution of estimates in the correlated case and a more narrowly spread distribution in the other case. The ranges along the x-axis are constant to illustrate the idea.

Thus, when the predictors in a linear model are highly correlated, the OLS approach can give imprecise estimates. The OLS estimates are good in the sense that they are unbiased estimators. Their pitfall, however, is that their variance is highly dependant on the correlations between predictors. The other modeling techniques described below find unique ways to address this problem.

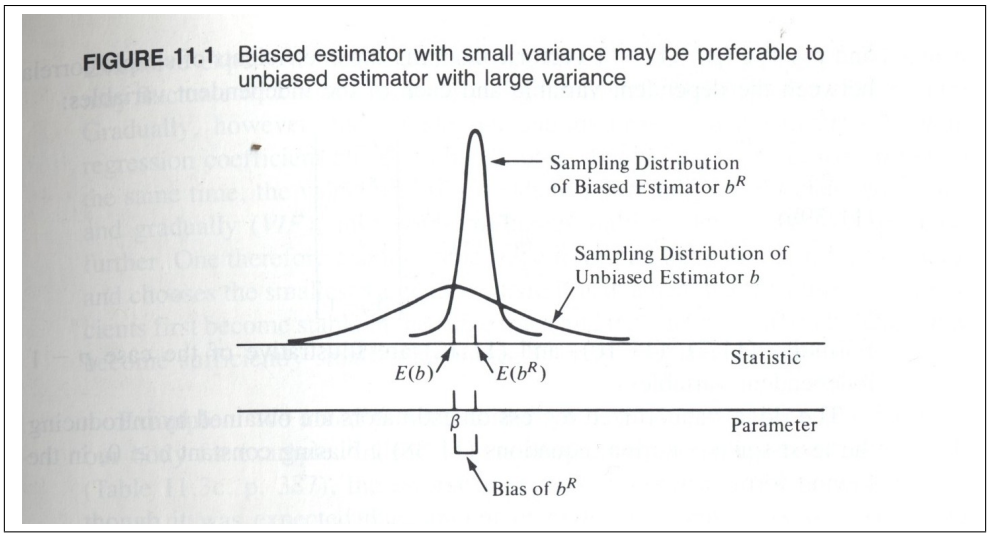

Figure 3: Biased estimators in Ridge Regression [4]

2.3 Performance Criterion: MSE

Good estimators should have low bias and low variance. To capture both of these, we look at the Mean Square Error (MSE) for an estimater $\hat{X}$ or a parameter X , defined as

$$E[(\hat{X} - X)^2].$$

Theorem 3. *The Mean Square Error of $\hat{X}$ is equal to the sum of the variance of $\hat{X}$ and the square of the bias of $\hat{X}$. Mathematically, $E[(\hat{X} - X)^2] = (\text{variance of } \hat{X}) + (\text{bias of } \hat{X})^2$.*

Proof.

$$\begin{aligned}
E[(\hat{X} - X)^2] &= E[\{(\hat{X} - E(\hat{X})) + (E(\hat{X}) - X)\}^2] \\
&= E[(\hat{X} - E(\hat{X}))^2] + E[(E(\hat{X}) - X)^2] + 2E[(\hat{X} - E(\hat{X}))(E(\hat{X}) - X))] \\
&= (\text{variance of } \hat{X}) + (\text{bias of } \hat{X})^2 + 2E[\hat{X}E(\hat{X}) - E[\hat{X}]^2 - \hat{X}X + E[\hat{X}]X) \\
&= (\text{variance of } \hat{X}) + (\text{bias of } \hat{X})^2 + 2(E[\hat{X}]^2 - E[\hat{X}]^2 - XE[\hat{X}] + XE[\hat{X}]) \\
&= (\text{variance of } \hat{X}) + (\text{bias of } \hat{X})^2 \qquad (3)
\end{aligned}$$

□

So, we often use MSE as the performance criterion when doing model selection. We saw that the bias of the OLS estimates is zero, but their variance can be high. Some modeling techniques introduce bias into their estimates in a way that lowers their variance and minimizes the MSE. See Figure 3 for an illustration of this idea.

2.4 Model Shrinkage

Including more predictors in our model will almost always improve the fit of the model. This kitchen sink approach, however, has huge pitfalls including but not limited to multiple collinear relationships between predictors. This causes the optimal $\hat{\beta}$ coefficients to be highly sensitive to small changes in X and largely unstable due to the matrix $X^T X$ being nearly singular (barely invertible). More recently we have developed more complex measures for predictive power.

Ideally, we want our model to emphasize the stronger predictors of the outcome variable and diminish or eliminate the other predictors. In some way, shape, or form, this is exactly what the techniques below attempt to do.

2.5 Subset Selection

Why do we want to screen out data which may be useful in the model? For one, large data sets are time consuming and computationally inefficient. Having a model with fewer independent variables is also beneficial because it is easier to analyze and interpret. Furthermore, highly correlated variables often add little predictive power, but do negatively affect the sampling variation of the β coefficients. For the reasons discussed, it is desirable to screen out independent variables. We first look for the ones which are either not fundamental to the problem, are subject to large errors, or effectively duplicate other variables. This process is called subset selection.

Its goal, then, is to find the 'best' subset of independent variables and while the criteria to measure this may be objective and mechanical, the process of screening ought not be such. With a set of P predictors, there are 2^{P-1} possible subsets of independent variables. We run OLS analysis on each of these subsets to find their respective MSE values and pick the subset of predictors which minimizes the MSE. Subset selection often generates models that perform very well, but the process takes a lot of computing time and energy, especially as P gets large. So, while subset selection often outperforms the other techniques, it is not efficient for larger models.

Forward Stepwise Regression is an offshoot of subset selection that begins with all of the predictors and strategically removes one at a time from the model until the remaining model has optimized its MSE. This is a more computationally feasible approach, even with large P.

As Kutner [4] notes, subjective judgment should be used to ensure that important predictors are not being thrown away. When considering a potential predictor we look to see whether its β coefficient is statistically significant. That is, we seek predictors with high t-scores where

$$t_i^* = \hat{\beta}_i / s(\hat{\beta}_i)$$

recalling that $s(\hat{\beta}_i)$ is the sampling variation of $\hat{\beta}_i$. So, predictors with larger values tend to have high statistical significance. A pivotal predictor may be ignored because of a narrow range of values, resulting in a statistically insignificant t-score. This is where judgement need be employed.

2.6 Ridge Regression

Ridge Regression introduces a penalty contraint on the size of the model. Specifically, Ridge Regression controls how large the sum of the squares of the estimated coefficients. That is, ridge finds $\hat{\beta}^R$ that minimizes

$$\sum_{j=1}^{n}(y_j - X_i^T\beta)^2 \text{ subject to } \sum_{j=1}^{p}(\beta_j)^2 \leq t.$$

The parameter $t \geq 0$ is a tuning parameter which controls the amount of shrinkage applied to the model. It can be shown that the ridge regression estimates satisfy

$$\hat{\beta}_i^R = \frac{1}{1+\gamma}\hat{\beta}_i^{OLS}$$

where γ depends on the choice of t as above.

2.7 Lasso and its Extentions

Lasso controls the size of the sum of the absolute value of the estimated coefficients. That is, it finds $\hat{\beta}^L$ which minimizes

$$\sum_{i=1}^{n}(y_i - X_i^T\hat{\beta})^2 \text{ subject to } \sum_{j=1}^{p}|\hat{\beta}_j| \leq t.$$

Again $t \geq 0$ is a tuning parameter which controls the amount of shrinkage applied to the model. Solving for the lasso coefficients gives:

$$\hat{\beta}_i^L = \text{sgn}(\hat{\beta}_i^{OLS})(|\hat{\beta}_i^{OLS}| - \gamma)^+$$

where $\gamma > 0$ depends on the choice of t above. The $^+$ notation means that $\hat{\beta}_i^L = 0$ for $|\hat{\beta}_i^{OLS}| < \gamma$

We can see the differences between Lasso and Ridge in figure 4 above. Both produce estimates which scale down the OLS coefficients. But while ridge regression scales the OLS coefficients continuously, Lasso sets smaller OLS coefficients to zero and thereby helps shrink and interpret the model. Therefore we note that Lasso is a nice combination of subset selection and Ridge regression: it shrinks the model both by eliminating some predictors (similar to subset selection) and by scaling the coefficients on the remaining predictors (similar to ridge).

Adaptive and Group Lasso Adaptive Lasso is an extension of Lasso which helps to account for differences in the magnitudes of estimated $\hat{\beta}$ coefficients. Adaptive Lasso minimizes

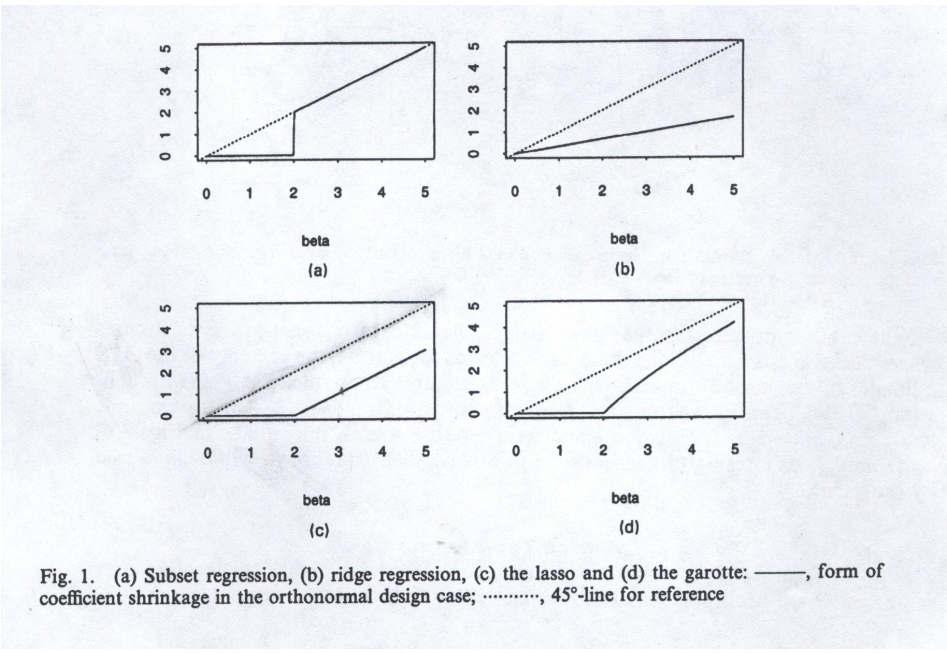

Fig. 1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garotte: ——, form of coefficient shrinkage in the orthonormal design case; ·····, 45°-line for reference

Figure 4: Model Shrinkage across techniques of Linear Modeling. The dotted line represents the OLS coefficients. [9]

$$\sum_{i=1}^{n}(y_i - X_i^T\beta^{aL})^2 \text{ subject to } \sum_{j=1}^{p}\frac{|\beta_j^{aL}|}{|\hat{\beta}_j^{OLS}|} \leq t$$

which increases the penalty for not eliminating predictors with smaller OLS coefficients and reduces the penalty for not scaling predictors with larger OLS coefficients. In this way, adaptive lasso has more of the advantages that subset selection has.

2.8 Choosing Optimal Contraints

AIC

For Lasso, choose the optimal tuning parameter (t constraint) we let t change across a range of values and choose the value that minimizes the AIC criteria according to:

$$AIC = \ln(\frac{1}{n}||Y - X\hat{\beta}||^2) + 2df/n. \qquad [10]$$

where *df* is equal to the number of non-zero coefficents in the model. We see that the first term measures the accuracy (bias) of the model while the second term captures the size (variance) of the model. Under this procedure, larger models are therefore penalized. In this way, minimizing the

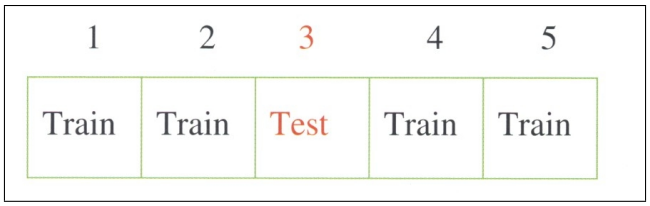

Figure 5: K fold cross validation, K=5 [2]

AIC statistic will drive more coefficients to zero in the selected model and will balance bias and variance.

For ridge regression, the process is more involved. AIC minimization is inappropriate for Ridge Regression. As discussed, the AIC procedure penalizes models that include more nonzero coefficients. Because Lasso sets some coefficients to zero, as t varies, so too will the number of nonzero estimated coefficients (df). Since AIC chooses the value of t that minimizes the AIC criterion, it will find a balance between subset selection and ridge regression.

Ridge Regression, on the other hand, uniformly scales the OLS coefficients but sets none equal to zero. This means that df is constant across all values of t. Thus, the AIC will be minimized for the smallest value in the range of t. In this case the Ridge solutions will be driven towards the OLS solutions because these minimize the bias of the model.

K-Fold Cross Validation

Instead, we use a technique called cross validation to optimize our t for ridge regression. Cross validation partitions the sample data into K equally sized complementary subsets. We remove one subset and fit the model using the data remaining in the $K-1$ subsets. Then we calculate the prediction error by using the fitted model to predict the data from the removed subset.

For an illustation, Figure 5 above shows data divided into 5 groups. For the ith part ($i = 1, 2, 3, 4, 5$), we fit the model to the other 4 parts of the data and calculate the prediction error of the model when predicting the ith part of the data. We repeat the process across all values of i and combine the values of prediction errors.

Then, we let the t constraint vary, repeating the above process, and choose the value of t which minimizes the total sum of the squared residual errors. This becomes our optimal contraint.

3 Asymptotics

3.1 Motivation

As shown in Figure 3, some models introduce bias into their estimators to lower variance and improve overall prediction accuracy. We know that the OLS estimates whose solutions minimize the sum of the squares of the residual errors produce unbiased estimators. By contraining the size

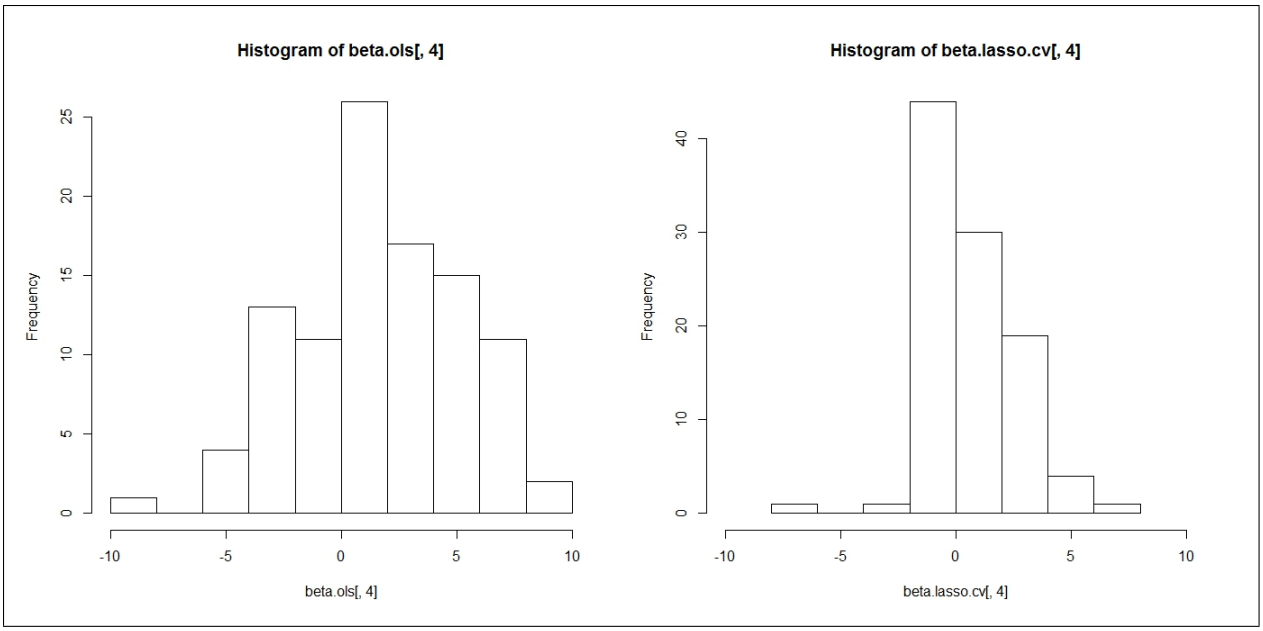

Figure 6: Sampling Variations in $\hat{\beta}^{OLS}$ and $\hat{\beta}^{L}$

of the model (as Lasso and Ridge Regression do), we introduce some bias into the estimates.

To illustrate this, we plotted the histograms of some OLS and Lasso estimates of $\beta_4 = 2$ of 100 simulations, each with 20 observations from a population

$$y = X\beta + 3\epsilon.$$

We had $\beta = (2, 2, 2, 2, 0, 0, 0, 0)^T$ and ϵ is the standard normal. There was a pairwise correlation between predictors x_i and x_j of $\rho^{|i-j|}$ with $\rho = .9$. The results are shown in Figure 6. Note that the OLS estimates are widely spread and roughly centered around 2. The Lasso estimates, on the other hand, are more narrowly spread but their center is closer to 0. This is a perfect recreation of Figure 3.

A well known result is that Lasso estimates are consistent. Essentially, they are biased with finite sample size ($n < \infty$) but their bias approaches zero as n approaches infinity. En route to proving this, we will first show that the OLS estimates are asymptotically normal.

3.2 Some Notation

First, we will extend our notion of a sequence of real numbers to a sequence of random variables. The motivation for doing so should become clear to the reader before proceeding. Much of statistics is focused on estimating parameters (such as the mean, variance, etc.) of a population of an unknown distribution. We do so by collecting random observations $X_1, X_2, ..., X_n$ (called random variables) from the population distribution and then use these to generate an estimate $\hat{\theta}_n = \hat{\theta}(X_1, ..., X_n)$ of a population parameter θ. It should be noted that the estimate $\hat{\theta}_n$ is itself a random variable due to its dependence on the set of random variables.

Specifically, we attempt to find the analogous concept of a limit for a sequence of random variables. For example, let X_n be the number of heads observed after flipping a fair coin n times. One may expect that

$$X_n/n \to 1/2$$

as $n \to \infty$. Upon closer inspection however, we find some glaring pitfalls with this notation: it is possible, albeit unlikely, that $X_n/n = 1$ even for large n. We can, however say that X_n/n is likely to be close to $1/2$. This motivates our definition of convergence in probablity and in distribution.

Definition 1. *A sequence of random variables X_n converges to a constant c in probability, denoted $X_n \xrightarrow{p} c$ if and only if for each $\epsilon > 0$*

$$P[|X_n - c| < \epsilon] \to 1$$

as $n \to \infty$. Equivalently, $P(|X_n - c| \geq \epsilon) \to 0$.

As Lehmann [6] notes, this condition may be hard to verify. We use the following lemma to find an equivalent condition for convergence in probability.

Lemma 1. Chebyshev Inequality. *For any random variable X_n and any constants $\epsilon > 0$ and c,*

$$E[(X_n - c)^2] \geq \epsilon^2 P[|X_n - c| \geq \epsilon]$$

The proof is omitted but we can now find an alternative condition for convergence in probablity.

Theorem 4. *A sufficient condition for $X_n \xrightarrow{p} c$ is that*

$$E[(X_n - c)^2] \to 0.$$

Proof. Suppose $E[(X_n - c)^2] \to 0$. Fix $\epsilon > 0$. By the previous lemma we find

$$P[|X_n - c| \geq \epsilon] \leq 1/\epsilon^2 E[(X_n - c)^2]$$

Letting $n \to \infty$ and noting the choice of ϵ is arbitrary gives the desired result. □

Since it is often much easier to show that $E(X_n - c)^2$ converges to zero, this theorem is very useful. We will discuss one last related topic (using suggestive notation for our parameters) before moving on.

Definition 2. *A sequence of estimators $\hat{X}_n$ of a parameter X is consistent if*

$$\hat{X}_n \xrightarrow{p} X.$$

Theorem 5. *A sufficient condition for $\hat{X}_n$ to be a consistent estimator of X is that both the bias and the variance of $\hat{X}_n$ tend to zero as $n \to \infty$.*

Proof. Theorem 3 along with Theorem 4 give the desired result. □

Theorem 1 shows that Mean Square Error, MSE $= E[(\hat{X}_n - X)^2]$ can be decomposed into terms having to with both the bias and the variance of X_n. Theorems 2 and 3 show that the limiting behavior of the MSE of X_n determines the consistency of X_n.

We introduce a few concepts to expand our asymptotic analysis of sequences before visiting the limiting distributions of the Lasso estimators.

Definition 3. *Two sequences* $\{a_n\}$ *and* $\{b_n\}$ *are asymptotically equivalent, denoted* $a_n \sim b_n$*, if* $a_n/b_n \to 1$ *as* $n \to \infty$*.*

In the case where b_n approaches a finite limit $b \neq 0$, we can say that a_n approaches the same limit. In the case that the limit b is 0 or $\pm\infty$, we must use more caution: $\{n\}$ and $\{n^2\}$ both approach ∞. They are not, however, equivalent since their ratio does not tend to 1. We thus need to look at order relations and rates of convergence.

Definition 4. *A sequence* $\{a_n\}$ *has order smaller than the sequence* $\{b_n\}$*, denoted* $a_n = o(b_n)$ *as* $n \to \infty$ *if* $a_n/b_n \to 0$*.*

Lemma 2. *A sequence* $\{a_n\}$ *satisfies* $a_n = o(1)$ *if and only if* $a_n \to 0$*.*

Definition 5. *Two sequences* $\{a_n\}$ *and* $\{b_n\}$ *are said to be of the same order, denoted* $a_n \asymp b_n$ *if* $a_n/b_n \neq 0$ *is finite. That is, there exists positive numbers m, M, and N such that*

$$m < |a_n/b_n| < M, \quad n > N.$$

Combining these gives rise to big O notation:

Definition 6. *A sequence* $\{a_n\}$ *has order smaller than or equal to the sequence* $\{b_n\}$*, denoted* $a_n = O(b_n)$ *as* $n \to \infty$ *if* a_n/b_n *is bounded.*

3.3 The Central Limit Theorem and the Delta Method

The Central Limit Theorem is one of the most widely cited theorem in statistics. It involves the asymptotic distribution of sample means and will be central to the analysis of behaviors of linear models in the remainder of this paper.

Theorem 6. Central Limit Theorem [6]. *Let* X_i*,* $i = 1, 2, ..., n$ *be independently and identically distributed from an unknown distribution with* $E(X_i) = \zeta$ *and* $Var(X_i) = \sigma^2 < \infty$*. Then,*

$$\sqrt{n}(\bar{X} - \zeta)/\sigma \xrightarrow{d} N(0, 1)$$

or equivalently,

$$\sqrt{n}(\bar{X} - \zeta) \xrightarrow{d} N(0, \sigma^2).$$

This theorem is incredibly powerful because no matter what population distribution the X_i come from, the distribution of their sample means converges in distribution to the normal distribution. Some other useful tools are Taylor's Theorem and the Delta Method.

Theorem 7. Taylor's Theorem [6]. *Suppose that $f(x)$ has r derivatives at a the point a. Then,*

$$f(a+\Delta) = f(a) + \Delta f'(a) + ... + \frac{\Delta^r}{r!} f^{(r)}(a) + o(\Delta^r).$$

Theorem 8. Delta Method [6]. *If*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \tau^2)$$

then,

$$\sqrt{n}[f(T_n) - f(\theta)] \xrightarrow{d} N(0, \tau^2[f'(\theta)]^2).$$

Proof. By Taylor's Theorem, with $a = \theta$ and $\Delta = T_n - \theta$,

$$f(T_n) = f(\theta) + (T_n - \theta)f'(\theta) + o_p(T_n - \theta)$$

and hence

$$\sqrt{n}[f(T_n) - f(\theta)] = \sqrt{n}(T_n - \theta)f'(\theta) + o_p[\sqrt{n}(T_n - \theta)].$$

The first term on the right hand side tends in distribution to $N(0, \tau^2[f'(\theta)]^2)$. On the other hand, we know that $\sqrt{n}(T_n - \theta)$ is bounded in probability and hence, the remainder tends to zero in probability. The result follows from previous results. □

3.4 Asympotic Normality of the LS Estimators

The paper by Knight and Fu [3] notes the apparently well known fact that the Ordinary Least Squares estimator $\hat{\beta}_n$ satisfies

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 C^{-1})$$

under the conditions

$$C_n = \frac{1}{n}\sum_{i=1}^{\infty}(\mathbf{x_i}\mathbf{x_i}^T) \to C$$

where C is a nonneagtive definite matrix and

$$\frac{1}{n}\max_{1\le i\le n} \mathbf{x_i}^T\mathbf{x_i} \to 0.$$

We will prove this here for the one dimensional ($p = 1$) case and note that the multivariate case is simlply a generalization shown using Slutzky's Theorem.

We have shown that $\hat{\beta}^{OLS} = (X^TX)^{-1}X^T\mathbf{y}$. In summation notation this translates to

$$\hat{\beta} = \Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}^T\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{y_i}\Big)$$

$$= \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \sum_{i=1}^{n} \frac{\mathbf{x_i}}{n} \end{bmatrix}^{-1} \begin{bmatrix} \bar{y} \\ \sum_{i=1}^{n} \frac{\mathbf{x_i y_i}}{n} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

(for the simple case where p=1) which, after inverting the matrix on the left and solving for $\hat{\beta}_1$ can be shown to give

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(\mathbf{x_i} - \bar{x})(\mathbf{y_i} - \bar{y})}{\sum_{i=1}^{n}(\mathbf{x_i} - \bar{x})^2} = \sum_{i=1}^{n} \mathbf{d_{ni} y_i} \qquad \mathbf{d_{ni}} = \frac{(\mathbf{x_i} - \bar{x})}{\sum_{i=1}^{n}(\mathbf{x_i} - \bar{x})^2}$$

We invoke the following theorem from Lehmann [6] to arrive at our desired result.

Theorem 9 [6]. *Let $Y_1, ..., Y_n$ be independently and identically distributed with $E(Y_i) = 0$ and $Var(Y_i) = \sigma^2 > 0$. Then*

$$\frac{\sum_{i=1}^{n} \mathbf{d_{ni} Y_i}}{\sigma \sqrt{\sum_{i=1}^{n} \mathbf{d_{ni}}^2}} \xrightarrow{d} N(0,1).$$

If we normalize our $\mathbf{y_i}$ to be $\mathbf{Y_i} = \mathbf{y_i} - \bar{y}$ then, since our $\hat{\beta}$ is an unbiased estimator of β, the conditions of the above theorem are met and we find that:

$$\sum_{i=1}^{n} \mathbf{d_{ni}}(\mathbf{y_i} - \bar{\mathbf{y}}) \Big(\frac{\sqrt{\sum_{i=1}^{n}(\mathbf{x_i} - \bar{x})^2}}{\sigma} \Big) = \frac{\sum_{i=1}^{n} \mathbf{d_{ni} Y_i}}{\sigma \sqrt{\sum_{i=1}^{n} \mathbf{d_{ni}}^2}} \xrightarrow{d} N(0,1).$$

Which is the univariate case of the result given in Knight [3].

3.5 Asymptotic Consistency of Lasso Estimates and its Limiting Distribution

Here, we again assume that the matrix C, defined above as

$$C_n = \frac{1}{n} \sum_{i=1}^{\infty} (\mathbf{x_i} \mathbf{x_i}^T) \to C$$

is nonsingular. We consider the asymptotic behavior of Lasso's objective function by first defining a random variable $Z_n(\hat{\beta})$:

$$Z_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^T\hat{\beta})^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j|$$

which, as we have seen, is minimized at $\hat{\beta} = \hat{\beta}_n^L$. The following theorem tells us that $\hat{\beta}_n^L$ is a consistent estimator of β provided that $\lambda_n = o(n)$

Theorem 10 [3]. *If C is nonsingular and $\lambda_n/n \to \lambda_0 \geq 0$ then $\hat{\beta}_n^L \xrightarrow{p} argmin(Z)$ where*

$$Z(\hat{\beta}) = (\hat{\beta} - \beta)^T C(\hat{\beta} - \beta) + \lambda_0 \sum_{j=1}^{p}|\hat{\beta}_j|$$

Thus, if $\lambda_n = o(n)$ then $\lambda_0 = 0$ and $argmin(Z) = \beta$ so that $\hat{\beta}_n^L$ is consistent.

Proof. We will show that $Z_n(\hat{\beta})$, defined above, converges in probability to $Z(\hat{\beta}) + \sigma^2$. The result will follow by applying established previous results from Pollard [7]. To show convergence of Z_n we translate it into matrix notation.

$$\begin{aligned}
Z_n(\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^T\hat{\beta})^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j| \\
&= \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j| \\
&= \frac{1}{n}((\epsilon + X\beta) - X\hat{\beta})^T((\epsilon + X\beta) - X\hat{\beta}) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j| \\
&= \frac{1}{n}(\epsilon + X(\beta - \hat{\beta}))^T(\epsilon + X(\beta - \hat{\beta})) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j| \\
&= \frac{1}{n}[\epsilon^T\epsilon + 2\epsilon X(\beta - \hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})] + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\hat{\beta}_j|
\end{aligned}$$

Now, we let $n \to \infty$ and note the following facts:

$$\sigma^2 I_{nxn} = var(\epsilon) = E(\epsilon^T\epsilon) + E(\epsilon)^T E(\epsilon) = E(\epsilon^T\epsilon).$$

$$\frac{1}{n}\epsilon^t\epsilon \xrightarrow{p} E[\epsilon^T\epsilon] = \sigma^2 I_{nxn} \text{ (by the law of large numbers)}$$

$$\frac{1}{n}X^TX \to C$$

$$E[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i] = \frac{1}{n}\sum_{i=1}^{n}E[\epsilon_i] = 0$$

$$2\frac{1}{n}\epsilon^T X(\beta - \hat{\beta}) \to 2E[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i]X^T(\beta - \hat{\beta}) = 0 \text{ (by the law of large numbers)}$$

$$\frac{\lambda_n}{n} \to \lambda_o = 0$$

so that

$$Z_n(\hat{\beta}) \xrightarrow{p} \sigma^2 + (\beta - \hat{\beta})^T C(\beta - \hat{\beta}) + \lambda_0 \sum_{j=1}^{p}|\hat{\beta}_j| = Z(\hat{\beta}) + \sigma^2$$

The pointwise convergence of $Z_n(\hat{\beta})$ to $Z(\hat{\beta}) + \sigma^2$ allows us to conclude that

$$\sup_{\hat{\beta}\in K}|Z_n(\hat{\beta}) - Z(\hat{\beta}) - \sigma^2| \xrightarrow{p} 0$$

for any compact set K by Pollard [7] and that

$$\hat{\beta}_n^L = O_p(1).$$

It follows that

$$argmin(Z_n) \xrightarrow{p} argmin(Z).$$

[3] □

This tells us that the lasso estimates are consistent. To find their limiting distribution, we invoke the following theorem.

Theorem 11 [3]. *If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and C is nonsingular then*

$$\sqrt{n}(\hat{\beta}_n^L - \beta) \xrightarrow{d} argmin(V)$$

where

$$V(\boldsymbol{u}) = \boldsymbol{u}^T C\boldsymbol{u} - 2\boldsymbol{u}^T\boldsymbol{W} + \lambda_0\sum_{j=1}^{p}u_j sgn(\hat{\beta}_j)$$

and $\boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 C)$. [3]

Proof. Define $V_n(\mathbf{u}) = \left(||\epsilon - X\mathbf{u}/\sqrt{n}||^2 + \lambda_n \sum_{j=1}^{p} |\beta_j + u_j/\sqrt{n}| \right) - \left(||\epsilon||^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \right)$ and note that $V_n(\mathbf{u}) = Q_n(\beta + u/\sqrt{n}) - Q_n(\beta)$ where

$$Q_n(z) = ||Y - Xz||^2 + \lambda_n \sum_{j=1}^{p} |z_j|$$

We find that $argmin\ V_n = \sqrt{n}(\hat{\beta}_n - \beta)$ by noting that

$$V_n(\sqrt{n}(\hat{\beta}_n - \beta)) = Q_n(\beta + (\hat{\beta}_n - \beta)) - Q_n(\beta) = Q_n(\hat{\beta}_n) + C$$

and that the objective function Q_n is analogous to the Lasso objective function, minimized at $\hat{\beta}_n$.

To show convergence of V_n, note that

$$\begin{aligned} Q_n(\beta + \mathbf{u}/\sqrt{n}) - Q_n(\beta) &= |(Y - X\beta) - X\mathbf{u}/\sqrt{n}|^T |(Y - X\beta) - X\mathbf{u}/\sqrt{n}| + \lambda_n \sum_{j=1}^{p} |\beta_j + u_j/\sqrt{n}| - Q_n(\beta) \\ &= ||\epsilon||^2 - 2\mathbf{u}^T/\sqrt{n} X^T + \mathbf{u}^T X^T X\mathbf{u}/n + \lambda_n \sum_{j=1}^{p} |\beta_j + u_j/\sqrt{n}| - (||\epsilon||^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|) \\ &= \mathbf{u}^T \frac{X^T X}{n} \mathbf{u} - 2\mathbf{u}^T \frac{X^T \epsilon}{\sqrt{n}} + \lambda_n \sum_{j=1}^{p} [|\beta_j + u_j/\sqrt{n}| - |\beta_j|] \end{aligned}$$

Then, letting $n \to \infty$ we apply the following facts:

$$\frac{1}{n} X^T X \to C$$

$$\frac{X^T \epsilon}{\sqrt{n}} \xrightarrow{d} \mathbf{W}$$

$$\frac{\lambda_n}{\sqrt{n}} \to \lambda_o$$

So that $V_n(\mathbf{u}) \xrightarrow{d} V(\mathbf{u})$, as defined above. □

4 Simulations in R

4.1 The Regression Scenario

For the following simulations in **R**, we are in the usual regression situation. We have collected n observational data points from a set of p predictors $\vec{x} = (x_1, x_2, ..., x_p)^T$ along with from a

response/outcome variable y. In this case we will model the data without the intercept term so we are left with the augmented data matrix:

$$\mathbf{X}|\ \mathbf{y} = \left[\begin{array}{cccccc|c} x_{11} & x_{21} & . & . & . & x_{p1} & y_1 \\ x_{12} & x_{22} & . & . & . & x_{p2} & y_2 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ x_{1n} & x_{2n} & . & . & . & x_{pn} & y_n \end{array}\right]$$

In the following simulations, we compare the Ordinary Least Squares Estimates with the estimates of Lasso, Ridge Regression, Adaptive Lasso, and Forward Stepwise Regression. To optimize the tuning parameters we used five fold cross validation in ridge regression and AIC minimization in Lasso and its extensions. We also used cross validation in Lasso for comparative purposes. The *steps* procedure in **R** was used to perform forward stepwise regression. These examples are consistent with those published in Tibshirani's paper *Regression Shrinkage and Selection via the Lasso* [9] in hopes of recreating his results.

4.2 Example 1

This example simulated 25 data sets, each with 20 observations from the population modeled as: $y = X\beta + \sigma\epsilon$ where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, ϵ is the standard normal, and $\sigma = 3$. There was a pairwise correlation between predictors x_i and x_j of $\rho^{|i-j|}$ with $\rho = .5$ Table 1 shows the median Mean Square Errors from each of the techniques. We use the median because it is a robust to outliers. We see that Lasso and Ridge Regression perform best. Table 2 shows the proportion of models from Lasso and forward stepwise regressions which include each β coefficient. We see that Ridge and Lasso perform comparably well and both outperform stepwise and OLS by a good measure. Table 2 shows that Lasso retains the correct coefficients more often than does forward stepwise. On average forward stepwise has closer to the correct number of zero coefficients (5) but suffers from high variablity in the nonzeo coefficients. Figure 7 below shows the inter quartile range of β estimates of the four linear models over the 25 data sets. We see that Lasso and forward stepwise are able to drive the zero β coefficents to zero very successfully.

Table 1: Results for Example 1

Method	Median MSE	Avg. no. of 0 coefficients
OLS	5.27	0
Ridge Regression	1.43	0
Lasso (AIC)	1.77	3.26
Lasso (cross-validation)	1.40	3.44
Adaptive Lasso	2.18	3.5
Forward Stepwise	2.45	3.62

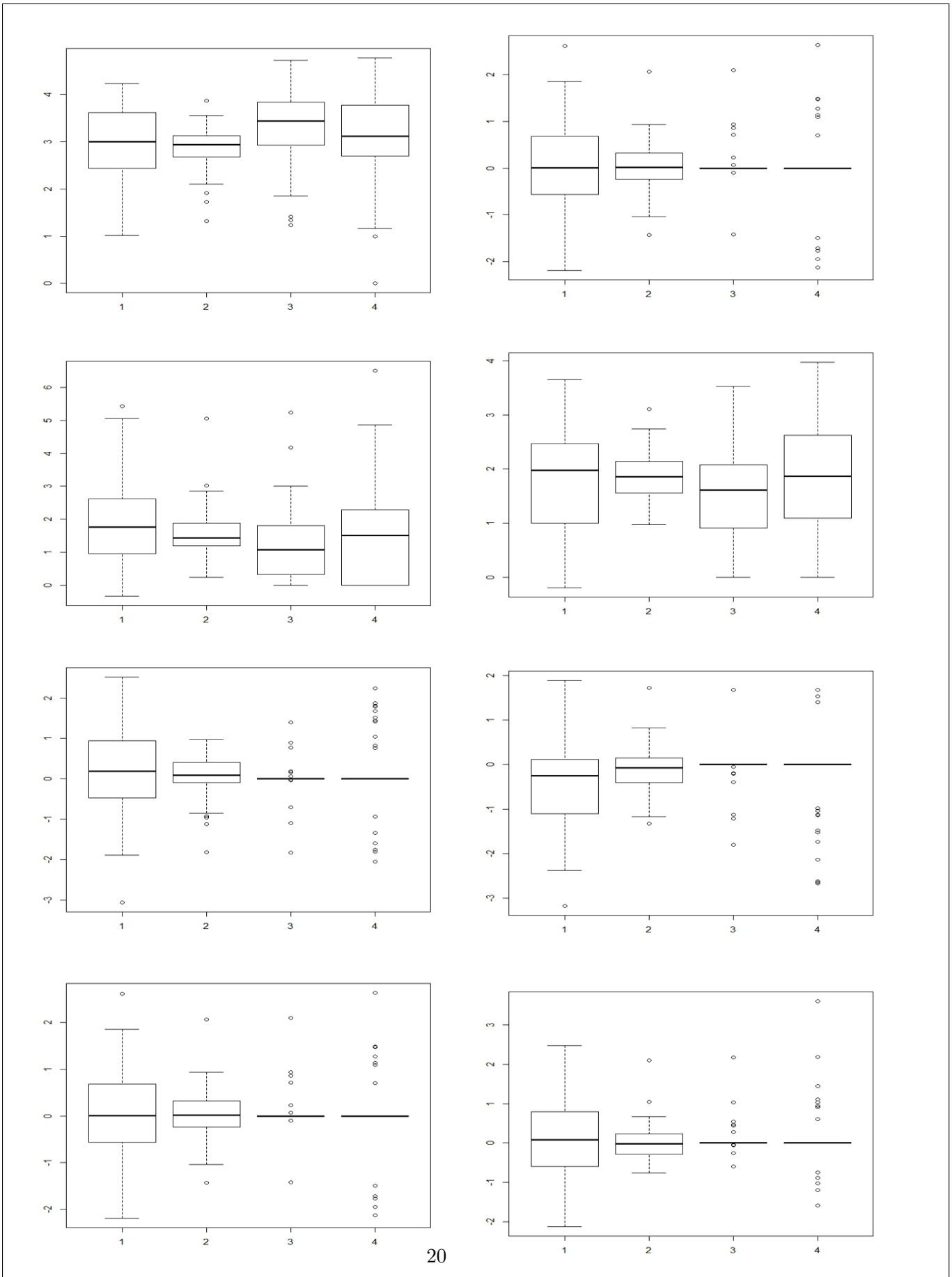

Figure 7: Estimates of the 8 coefficients in example 1. 1=OLS 2=Ridge 3=Lasso 4=Fwd Stepwise

Table 2: Proportion of Models including β coefficients

Coefficient	Lasso	Forward Stepwise
β_1=3	1.0	.98
β_2=1.5	.82	.78
β_3=0	.22	.34
β_4=0	.22	.26
β_5=2	.98	.88
β_6=0	.16	.30
β_7=0	.22	.26
β_8=0	.24	.18

4.3 Example 2: Highly Correlated Predictors

This example is similar to example 1 except we set $\rho = .9$ to increase the correlation between predictors. Table 3 shows the results and we see that ridge regression, Lasso and its extensions are able to handle the correlation issue by significantly outperforming the OLS estimates. Again we see that forward stepwise, on average, chooses the correct number of zero coefficients. Here, the variance problem in the nonzero coefficients in magnified, as its median mean square error is well above that of lasso and ridge regression.

Table 3: Results for Example 2

Method	Median MSE	Avg. no. of 0 coefficients
OLS	4.35	0
Ridge	.27	0
Lasso (AIC)	.64	5.76
Lasso (cross-validation)	.75	5.2
Adaptive Lasso	1.92	4.92
Forward Stepwise	2.99	5.08

4.4 Example 3: Sparse Signals

Here we let $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$ and $\sigma = 2$. As before, there are eight coefficients to estimate but only the first one is non-zero. In this sense the signals of β are sparse. This set up should be well suited for Lasso and its extensions. The results in Table 4 confirm this: Lasso's median mean square error is unprecedentedly low.

4.5 Example 4: Smaller Signals

Here we had the same set-up as in example 1 except $\beta_j = .85$ for all $1 \leq j \leq 8$. That is, the signals of β are equally weighted but relativiely smaller in magnitude. The results in Table 5 reveals that Ridge Regression outperforms the others. Lasso comes in second. OLS actually outperforms

Table 4: Results for Example 3

Method	Median MSE	Avg. no. of 0 coefficients
OLS	4.71	0
Ridge	1.07	0
Lasso (AIC)	.26	5.44
Lasso (cross-validation)	.018	5.68
Adaptive Lasso	1.18	5.12
Forward Stepwise	2.41	5.12

forward stepwise selection in this case. This could be due to the fact that model selection is useless in this case. Since all β elements are equal, it is impossible to forward stepwise to choose the strongest signals.

Table 5: Results for Example 4

Method	Median MSE	Avg. no. of 0 coefficients
OLS	6.6	0
Ridge	1.7	0
Lasso (AIC)	3.7	2.4
Lasso (cross-validation)	3.4	2.16
Adaptive Lasso	5.89	3.64
Forward Stepwise	6.93	4

4.6 Example 5: Larger Models

Here, we are set up to analyze the performance of the estimators in larger models to better reflect a real world application. We simulated 25 data sets, each with 100 observations and 40 predictors. The coefficient vector was $\beta = (0, 0, ..., 0, 2, 2, ..., 2, 0, 0, ..., 0, 2, 2, ..., 2)^T$ and $\sigma = 15$. As in example 1, we had $\rho = .5$. Table 6 shows the performance results. We see that ridge outperforms the others but is doing no model selection. Lasso under cross validation has a reasonably low MSE and does a great job at model selection.

Table 6: Results for Example 5

Method	Median MSE	Avg. no. of 0 coefficients
OLS	164.99	0
Ridge	10.81	0
Lasso (AIC)	111.83	33.08
Lasso (cross-validation)	43.45	21.2
Adaptive Lasso	95.91	20.72
Forward Stepwise	305.22	27.4

5 Conclusion

We began our discussion by studying the Ordinary Least Squares technique. After some sensitivity analysis where we increased the correlation between predictors, we found that the OLS estimates are unreliable estimators when there is collinearity between predictors. We then touched on alternative modeling techniques that attempt to improve upon OLS and remedy its shortcomings. By constraining the size of the model, we introduce bias into our estimates in an attempt to lower the overall mean square error of the model. By applying the Central Limit Theorem and Taylor's Theorem, we found that the distribution of OLS estimates are centered around the true parameter value and approach normality as n approaches infitity. After some legwork, we also found that the Lasso estimates are asymptotically consistent. That is, as n approaches infinity, the bias in the Lasso model approaches zero. Further, the distribution of the Lasso estimates is asymptotically normal. This is reassuring because with a lagre sample size, we are confident that the bias of the Lasso estimates is negligible.

To test performance of the various models with finite sample size we ran simulations in **R**. For the most part, our mathematical intuitions were correct. In small models with sparse signals, Lasso outperformed all other techniques by a significant margin. Ridge was next followed by forward stepwise. In models with more signals and larger signals, Here Lasso and Ridge performed well (especially with highly correlated predictors) while forward stepwise and OLS did poorly. Finally, in models with a large number of smaller signals (no zero coefficients), Ridge Regression does best by far. Lasso still outperforms both OLS and forward stepwise. Overall, when the true model is unknown, we can conclude that Lasso will do the best job at balancing model selection with prediction accuracy.

By using improved linear modeling techniques such as subset selection, ridge regression and Lasso, we are able to efficiently parse though large data sets (full of redundant and/or useless information) and accurately predict outcomes. These tools will continue to be improved upon and applied to larger and more complicated models.

References

[1] Gleser et. al. *The Limiting Distribution of Least Squares in Errors-in-Variables Regression Model* , The Annals of Statistics, Mar. 1987.

[2] Hastie, et. al. *The Elements of Statistical Learning*, Springer, 2009.

[3] Knight, K. and Fu, W. *Asympotics for Lasso-Type Estimators*, University of Toronto, 2000.

[4] Kutner, et. al. *Applied Linear Statistical Models, 5th ed.*, McGraw-Hill, 2004.

[5] Lay, David C. Linear Algebra and its Applications Pearson, 2006.

[6] Lehmann, E.L. *Elements of Large-Sample Theory*, Springer, 1999.

[7] Pollard, David *Asymptotics for Least Absolute Deviation Regression Estimators*, Yale University, 1991.

[8] Srivastava, M. S. *On Fixed-width Confidence Bounds for Regression Parameters and Mean Vector*, Princeton University, 1967.

[9] Tibshirani, Robert. *Regression Shrinkage and Selection via the Lasso*, University of Toronto, 1994.

[10] Wang, H and Leng, C. *A note on adaptive group lasso*, Peking University, 2008.