Abstract

The goal of this paper is to do some basic proofs for lasso and have a deep understanding of linear regression. In this paper, firstly I give a review of methods in linear regression, and most concerns with the method of lasso. Lasso for ‘least absolute shrinkage and selection operator’ is a regularized version of method adds a constraint which uses L^1 norm less or equal to a given value t. By doing so, some predictor coefficients would be shrank and some others might be set to 0. We can attain good interpretation and prediction accuracy by using lasso method. Secondly, I provide some basic proofs for lasso, which would be very helpful in understanding lasso. Additionally, some geometric graphs are also given and one example is illustrated.

Acknowledgements

Firstly, I would like to sincerely express my gratitude to my advisor, Professor Wu, for his patient discussion with me on this project and his nice teaching and help in class.

I would also like to say thank you to my Professor Petruccelli, Professor Nandram, Professor Doytchinova, and Professor Kim at Worcester Polytechnic Institute for all your kindness and teaching to me.

Thirdly, I need to thank my parents for their past 25 years supporting and loving. I feel really blissful to be with you.

Contents

Chapter 1

1.1 Introduction

In nowadays data analysis, the method of linear regression has been very popular. There are several methods and algorithms been developed these years. The most familiar one that is often used is least square estimation. It is used more extensively than other estimation procedure for building regression models and was exclusively used prior to the 1970s. It is a model where the sum of squared residuals has its least value. The motivation is that the data points will be "close" to the fitted regression line if the errors in fit are rendered small. [5] A residual here means the difference between a real observed data value and the value predicted by the model being used. As we know, in biostatistics or social sciences, sometimes, there are thousands of underlying predictors, which makes the linear model hard to interpret and communicate or may even experience the risk of over-fitting. What's more, the average variability of predicted response is $p\sigma^2 / n$, then, as a result, the large models would produce more statistically variable predictions. [1]

Considering the prediction accuracy and interpretability, several other methods have been introduced, such as ridge regression and subset regression. Ridge regression is one exciting research topics during 1970s and 1980s. Its popularity rose dramatically with the publication of the article by Hoerl and Kennard in Technometrics. It is a parameter estimation method to address the collinearity problem frequently arising in multiple linear regressions. The ridge regression methodology falls into the category of biased estimation techniques. It yields a class of biased estimators indexed by a scalar nonnegative parameter. The challenge is to determine which estimator within this class to use in the context of specific problem, namely, determining a best choice for the ridge parameter. [4] In this paper, I most concerned about the method called LASSO for ‘least absolute shrinkage and selection operator’, proposed by Professor Robert Tibshirani in the year 1996. This regularized version of method adds a constraint which uses L^1 norm less or equal to a given value t. By using this constraint, some predictor coefficients would be shrank and some others would be set to 0. Here the given value t will cause solution shrinkage to 0 and some coefficients may be exactly 0 if t is the less than the full least squares estimates $t_0 = \sum \hat{\beta}_j^0$. [1] The given value t can be fixed or got by bootstrap sample and then optimize it.

Chapter 2

Review of Related Literature

2.1 Linear Regression Model

The model $Y = X\beta + \varepsilon, \varepsilon \sim N_n(0, \sigma^2 I)$

Namely, $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-1,2} \\ & \vdots & & & \vdots \\ 1 & x_{1n} & x_{2n} & & x_{p-1,n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

A linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is approximately linear, and that the design matrix X has full column rank p. ε is the model error and each ε_i is assumed uncorrelated from observation to observation, with mean zero and constant variance. What's more, each x_{ij} is assumed fixed and are measured with negligible error.

A linear model is defined as a model that is linear in the parameters, namely linear in the coefficients. The relation between the response variable y and predictors x can be, for example, polynomial in nature, yet the model would still be a linear model. In some regression applications we need to do the transformations on the predictor variables. Natural log transformation and power transformation are often performed. The response variable y is often involved in the transformations.

2.2 The Least squares estimation

The least squares estimator $\hat{\beta}^{LSE}$ for regression coefficients in β is the vector that satisfies

$$\frac{\partial}{\partial\beta}\left[\left(y - X\overset{\wedge}{\beta}^{LSE}\right)^T\left(y - X\overset{\wedge}{\beta}^{LSE}\right)\right] = 0$$

Here, the $(y - X\beta)^T(y - X\beta)$ represents the residual sum of squares. Performing the partial derivative

$$-2X^T y + 2X^T X\overset{\wedge}{\beta}^{LSE} = 0$$

$$X^T X\overset{\wedge}{\beta}^{LSE} = X^T y$$

So, the least squares estimation solution is

$$\overset{\wedge}{\beta}^{LSE} = (X^T X)^{-1} X^T Y$$

This estimation procedure is a good one if $X^T X$, when in the form of correlation matrix, is nearly a unit matrix. However, if $X^T X$ is not nearly a unit matrix, the least squares estimates are sensitive to a number of “errors”. Estimation based on the matrix$\left[X^T X + kI_p\right], k \geq 0$ rather than on $X^T X$ has been used to found to be a procedure that can be used to help circumvent many of the difficulties associated with the ordinary least squares estimates.[3]

2.3 Positive definite matrix

Suppose that we have a column vector z (n elements) and $n \times n$ symmetric matrix A with typical element a_{ij} . Then the scalar quantity

$$z^T Az = \sum_{i=1}^{n} a_{ii} z_i^2 + 2\sum_{i=1}^{n}\sum_{j=1, j>i}^{n} a_{ij} z_i z_j$$

is called a quadratic form in z with matrix A.

A positive definite quadratic form is on that is greater than zero for all $z \neq 0$. A positive definite matrix A is one for which $z^T Az > 0$ for all $z \neq 0$. A positive semi-definite matrix is one for which $z^T Az \geq 0$ for all z, but $z^T Az = 0$ for some $z \neq 0$.

2.4 Ridge regression

For the case of collinearity, the diagonals do not dominate. This non-dominance of the diagonals causes at least one eigenvalue to be small. To make $X^T X$ behave more like the orthogonal case, we can increase the eigenvalues, decrease the determinant of the matrix, and hence decrease the elements of the inverse.

Suppose we consider replacing the matrix $X^T X$ by the matrix $\left(X^T X + kI\right)$, where k is a small positive quantity. The ridge regression estimator is found by solving for β_R in the system of equations

$$\left(X^T X + kI\right)\beta_R = X^T y$$

where $k \geq 0$ is often referred to as a shrinkage parameter. [5] The solution if given by

$$\beta_R = \left(X^T X + kI\right)^{-1} X^T y$$

Ridge regression minimizes

$$\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_j \beta_j^2$$

Or equivalently, minimizes

$$\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 \quad \textit{subject to} \sum_j \beta_j^2 \leq t.$$

The ridge solutions are

$$\frac{1}{1+\gamma}\hat{\beta}_j^{LSE}$$

where γ depends on λ or t.

2.5 Stepwise algorithm

1. Start with a “minimal model” that includes any covariates that must be in the model based on subject matter knowledge. A model with only an intercept is a possibility.
2. Stepwise addition
   (a) Add to the model the covariate that is not currently in the model and whose estimated parameter would have the largest t-statistic in absolute value. Call this statistic t_{enter}. Below, we would describe how to compute the t_{enter} statistics using updates to the current fit rather than completely new fits.
   (b) Fit the model including the new covariate using least squares and record the Generalized Cross Validation value of the fit.
3. Stepwise deletion

   Repeat until only the minimal model remains:
   (a) Delete from the current set of covariates (not including those in the minimal model) the covariate that has the smallest absolute t-statistic (t_{exit}).
   (b) Fit the model with the remaining covariates using least squares and record the GCV value of the fit.

   The final model is the one that has the smallest GCV value.

Let X be the design matrix containing column vectors of the variables that are currently in the model and H be the associated “hat” matrix:

$$H = X\left(X^T X\right)^{-1} X^T$$

Let I be an n by n identity matrix. Denote the j th element of a vector by $\{.\}_j$ and the j, j th entry of a matrix by $\{.\}_{j,j}$. Let $\hat{\sigma}$ be the residual standard deviation from a fit with all of the covariates in the model. Then,

$$t_{exit} = \left| \frac{\left\{ \left(X^T X \right)^{-1} X^T y \right\}_j}{\hat{\sigma} \sqrt{\left\{ \left(X^T X \right)^{-1} \right\}_{j,j}}} \right|$$

Suppose that x_k is the vector of covariates that is the candidate to enter. [2] Then,

$$t_{enter} = \left| \frac{\hat{\beta}_k}{se\left(\hat{\beta}_k \right)} \right| = \left| \frac{X^T \left(1 - H\right) y}{\hat{\sigma} \sqrt{X^T \left(1 - H\right) x_k}} \right|$$

2.6 Best subset algorithm

1. Start with a model selected as “best” by the Stepwise algorithm and suppose this model has uses $p_c \le p$ covariates.
2. Fit all the $\binom{p}{p_c}$ possible models with p_c covariates and record the GCV statistic for each.

The final model is the one that has the smallest GCV value. Variations on this algorithm consider all models of size $p_c - 1$ and $p_c + 1$ too, etc. [2]

Best Subsets Regression is a method used to help determine which predictor (independent) variables should be included in a multiple regression model. This method involves examining all of the models created from all possible combination of predictor variables. [2] Usually, we use prefer to use a statistical software program to do Best Subsets Regression.

2.7 Lasso

Let $\left(x^i, y_i\right)$, i=1, 2, …, N, where $x^i = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)^T$ are the predictors and y_i are the responses. We assume that the observations are independent or y_i s are conditionally independent given x_{ij} s. It's also supposed that x_{ij} are standardized so that $\sum_i x_{ij} / N = 0$, $\sum_i x_{ij}{}^2 / N = 1$. Let $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p\right)^T$, the lasso estimate $\left(\hat{\alpha}, \hat{\beta}\right)$ is defined by [1]

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\} \quad \textit{subjuct to} \sum\left|\beta_j\right| \le t. \tag{1}$$

Here, the I can show that the solution for α is $\hat{\alpha} = \bar{y}$.

Proof:

Compute the derivative of (1) and set it to 0, then we get:

$$2\sum_i\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right) = 0$$

$$\sum_i y_i - N\alpha - \sum_i \sum_j \beta_j x_{ij} = 0$$

$$\sum_i \sum_j \beta_j x_{ij} = \sum_j \beta_j \sum_i x_{ij} = 0$$

$$N\bar{y} - N\alpha - 0 = 0$$

$$\alpha = \bar{y}.$$

So, here, we can assume without loss of generality that $\bar{y} = 0$ and hence accordingly omit α. Here $t \ge 0$ is a tuning parameter and controls the amount of shrinkage which is applied t the estimates.

Then the simplified version of (1) can be written as:

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\} \quad \textit{subjuct to} \sum\left|\beta_j\right| \le t$$

2.8 Stationary point

A stationary point is an input to a function where the derivative is zero, where the function "stops" increasing or decreasing.

For the graph of a one-dimensional function, this corresponds to a point on the graph where the tangent is parallel to the x-axis. For the graph of a two-dimensional function, this corresponds to a point on the graph where the tangent plane is parallel to the x-y plane.

Stationary points in higher dimensions are usually referred to as critical points. Yet, Critical point is a more general definition: a critical point is either a stationary point or a point where the derivative is not defined. [6]

Chapter 3

Theorem

3.1Lagrange multipliers

In mathematical optimization, Lagrange multipliers method provides a strategy for finding the maximum or minimum of a function subject to constraints. For example, if we want to maximize f(x, y) subject to g(x, y) = c, then we introduce a new multiplier λ a called Lagrange multiplier, and define the Lagrange function as:

$$L(x,y,\lambda)=f(x,y)+\lambda\big(g(x,y)-c\big).$$

If there a point (x, y) is the maximum point for the original constraints，there exists a λ such that (x, y, λ) is a stationary point for the Lagrange function. However, not all stationary points yield a solution of the original problem. As a result, the method of Lagrange multipliers only provides a necessary condition for optimality in constrained problems.

A more general case, denote the objective function by f(x) and let the constraints be given by $g_k(x)$. Here, x is a vector. The domain of f should be an open set containing all points satisfying the constraints. Furthermore, f and the $g_k(x)$ must have continuous first partial derivatives and the gradients of the $g_k(x)$ must not be zero on the domain. Now, we can define the Lagrange function as:

$$L(x,\lambda)=f(x)+\sum_k\big(\lambda_k g_k(x)\big).$$

λ is a vector with independent elements λ_k.

Observe that both the optimization criteria and constraints $g_k(x)$ are compactly encoded as stationary points of the Lagrangian:

$$\begin{cases} \nabla_x L = 0 \\ \nabla_\lambda L = 0 \end{cases}$$

Implies that

$$\begin{cases} \nabla_x f = -\sum_k \left(\lambda_k \nabla_x g_k \right) \\ \qquad g_k(x) = 0 \end{cases}$$

Collectively, we can write above as:

$$\nabla L = 0 .$$

Then, we can solve a number of equations totaling the length of x plus the length of λ .

3.2 Karush-Kuhn-Tucker Conditions

The Karush–Kuhn–Tucker conditions, also known as the Kuhn-Tucker conditions, are necessary for a solution in nonlinear programming to be optimal, provided some regularity conditions are satisfied. It is a generalization of the method of Lagrange multipliers to inequality constraints. The conditions are named after William Karush, Harold W. Kuhn, and Albert W. Tucker.

Let us consider the following nonlinear optimization problem:

Minimize: $f(x)$

Subject to: $\begin{cases} g_i(x) \le 0 \\ h_j(x) = 0 \end{cases}$

where $g_i(x)$ (i=1, … , m) are the inequality constraints and are $h_j(x)$ (j=1, …, l) equality constraints. x is a vector.

Suppose that the objective function $f(x)$ is $f : R^n \to R$ and the constraint functions are $g_i : R^n \to R$ and $h_j : R^n \to R.$ Further, suppose they are continuously differentiable at a point x^* . If x^* is a local minimum that satisfies some regularity conditions, then there exist constraints μ_i (i=1, … , m) and λ_j (j=1, … , l) such that

（1） Stationary $\nabla f\left(x^*\right)+\sum_{i=1}^{m}\mu_i\nabla g_i\left(x^*\right)+\sum_{j=1}^{l}\lambda_j\nabla h_j\left(x^*\right)=0$

（2） $\begin{cases} g_i\left(x^*\right)\le 0, i=1,...,m \\ h_j\left(x^*\right)=0,\ j=1,...,l \end{cases}$

（3） $\mu_i \ge 0, i=1,...,m$

（4） Complementary slackness $\mu_i g_i\left(x^*\right)=0, i=1,...,m.$

Chapter 4

Special case of Lasso

4.1Orthonormal Design Case

Let X be $n \times p$ design matrix with $ij\,th$ entry x_{ij} , and suppose that $X^T X = I$. Then solutions to equation (1) are:

$$\hat{\beta}_j = sign\left(\hat{\beta}_j^0\right)\left(\left|\hat{\beta}_j^0\right| - \gamma\right)^+ \text{, where } \gamma = \left(\frac{\sum_j \left|\hat{\beta}_j^0\right| - t}{p}\right)^+ \qquad (3)$$

We can use Lagrange multipliers method and KKT (Karush–Kuhn–Tucker conditions) conditions to solve it.

Proof:

$$\min \sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 = \min \left\| Y - X\beta \right\|_2^2 , \left\|\beta\right\|_1 \le t$$

Since X is orthonormal, then

$$X^T X = I \Rightarrow \left(X^T X\right)^T = I$$

$$\arg\min_\beta \left\|Y - X\beta\right\|_2^2 = \arg\min_\beta \left\|X^T\left(Y - X\beta\right)\right\|_2^2 = \arg\min_\beta \left\|\left(X^T X\right)^T X^T\left(Y - X\beta\right)\right\|_2^2$$

$$= \arg\min_\beta \left\|\left(X^T X\right)^T X^T Y - \left(X^T X\right)^T X^T X\beta\right\|_2^2$$

$$= \arg\min_\beta \left\|\hat{\beta}^0 - \beta\right\|_2^2 , \, where\ \hat{\beta}^0 = \hat{\beta}^{LSE}$$

$$Then, let\ f\left(x\right) = \left\|\beta - \hat{\beta}^0\right\|_2^2 , g\left(x\right) = \sum_j \left|\beta_j\right| - t \le 0$$

$$L\left(x\right) = \left\|\beta - \beta^0\right\|_2^2 + \lambda\left(\sum_j \left|\beta_j\right| - t\right), \lambda \ge 0$$

$$\frac{\partial L}{\partial \beta_j} = 2(\beta_j - \beta_j^0) + \lambda sign(\beta_j) = 0$$

$$\Rightarrow \beta_j = \hat{\beta}_j^0 - \frac{\lambda}{2} sign\left(\beta_j\right) \quad （4）$$

$$\frac{\partial L}{\partial \lambda} = \sum_j \left|\beta_j\right| - t = 0$$

$$\Rightarrow \sum_j \left|\beta_j\right| = t \quad （5）$$

From (4), we have

$$\left(\hat{\beta}_j^0 - \beta_j\right) = \frac{\lambda}{2} sign\left(\beta_j\right)$$

$$\hat{\beta}_j^0 \, sign(\beta_j) - \left|\beta_j\right| = \lambda / 2 \ge 0$$

$$\Rightarrow \hat{\beta}_j^0 \, sign(\beta_j) \ge 0$$

$$\Rightarrow sign\left(\hat{\beta}_j^0\right) = sign\left(\beta_j\right)$$

$$Then, we\ have \left|\hat{\beta}_j^0\right| - \left|\beta_j\right| = \frac{\lambda}{2}$$

$$\sum_j \left|\hat{\beta}_j^0\right| - \sum_j \left|\beta_j\right| = p * \frac{\lambda}{2}$$

$$\frac{\sum_j \left|\hat{\beta}_j^0\right| - t}{p} = \frac{\lambda}{2} \ge 0$$

$$\left(\frac{\sum_j \left|\hat{\beta}_j^0\right| - t}{p}\right)^+ = \frac{\lambda}{2}$$

$$\beta_j = \overset{\wedge}{\beta}{}_j^0 - \left(\frac{\sum_j \left| \overset{\wedge}{\beta}{}_j^0 \right| - t}{p} \right)^+ sign\left(\beta_j\right)$$

$$= sign\left(\overset{\wedge}{\beta}{}_j^0 \right) \left(\overset{\wedge}{\beta}{}_j^0 \, sign(\overset{\wedge}{\beta}{}_j^0) - \left(\frac{\sum_j \left| \overset{\wedge}{\beta}{}_j^0 \right| - t}{p} \right)^+ \right)$$

$$= sign\left(\overset{\wedge}{\beta}{}_j^0 \right) \left(\left| \overset{\wedge}{\beta}{}_j^0 \right| - \left(\frac{\sum_j \left| \overset{\wedge}{\beta}{}_j^0 \right| - t}{p} \right)^+ \right)$$

Another condition that needs to be satisfied is that $\lambda * (\sum_j \left|\beta_j\right| - t) = 0$

(i)

If $\lambda = 0 \Rightarrow \left| \overset{\wedge}{\beta}{}_j^0 \right| - \left|\beta_j\right| = \frac{\lambda}{2} = 0 \Rightarrow \left| \overset{\wedge}{\beta}{}_j^0 \right| = \left|\beta_j\right| \Rightarrow \beta_j = \overset{\wedge}{\beta}{}_j^0$, j=1, 2,…p

It means that the solution is just OLS solution, no shrinkage.

(ii)

We need the solution satisfy that $\sum_j \left|\beta_j\right| - t = 0$

$$\sum_j \left| \left| \overset{\wedge}{\beta}{}_j^0 \right| - \left(\frac{\sum_j \left| \overset{\wedge}{\beta}{}_j^0 \right| - t}{p} \right)^+ \right| - t = 0$$

$$\Rightarrow \sum_j \left| \overset{\wedge}{\beta}{}_j^0 - \left(\frac{\sum_j \left| \overset{\wedge}{\beta}{}_j^0 \right| - t}{p} \right)^+ \right| - t = 0 \qquad (6)$$

$$\geq \sum_j \left(\hat{\beta}_j^{0} - \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+} \right) - t$$

$$= \sum_j \left| \hat{\beta}_j^{0} \right| - \sum_j \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+} - t$$

$$= \sum_j \left| \hat{\beta}_j^{0} \right| - p \times \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+} - t = 0$$

Only when $\left| \left| \hat{\beta}_j^{0} \right| - \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+} \right| = \left| \hat{\beta}_j^{0} \right| - \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+}$, $j = 1, 2, \ldots p$, (6) is satisfied. So,

$$\left| \hat{\beta}_j^{0} \right| - \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+} , j = 1, 2, \ldots p \, .$$

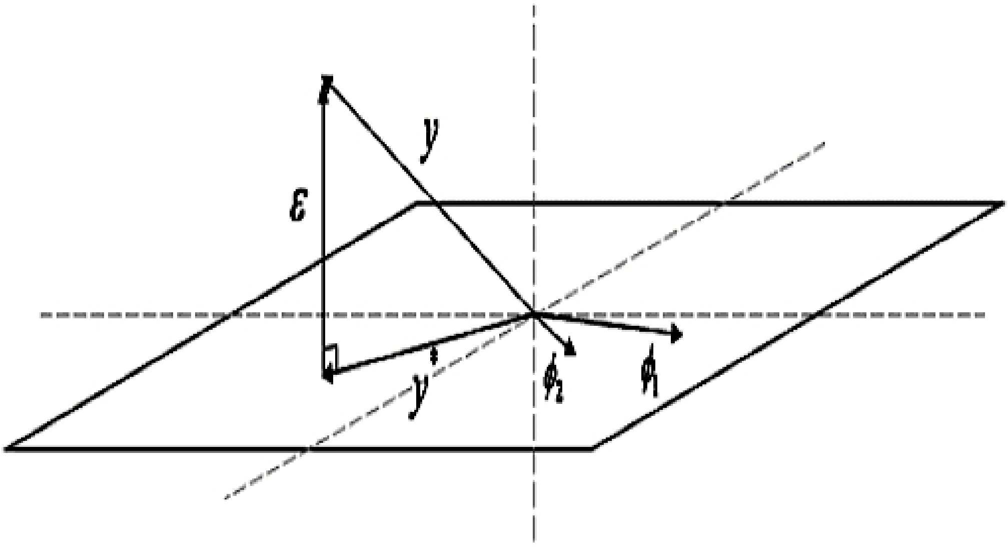
So, the solution is $\hat{\beta}_j = sign\left(\hat{\beta}_j^{0} \right) \left(\left| \hat{\beta}_j^{0} \right| - \gamma \right)^{+}$, where $\gamma = \left(\frac{\sum_j \left| \hat{\beta}_j^{0} \right| - t}{p} \right)^{+}$.

Chapter 5

Geometry

5.1 Geometry of Least squares

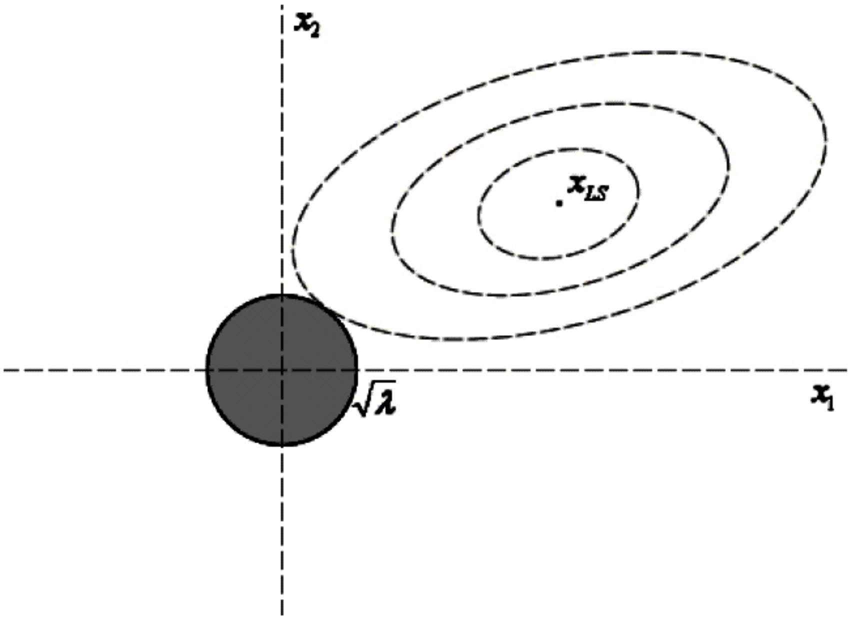
Figure 1 Estimation picture for least squares



Consider the picture given in Figure 1, which depicts a regression with n=3 observations and p=2 parameters. The three-dimensional axis system shown is in the y-observation space, and the vector y represents the observation vector in the observation space. The two-dimensional plane in the figure represents the estimation space. As we know, $X(X^TX)^{-1}X^T$ is the perpendicular projection onto the column space of X. Thus, $X(X^TX)^{-1}X^Ty = X\hat{\beta}^{LSE} = \hat{y}$ is perpendicular projection of y in the two-dimensional plane. We can understand it in another way. What point in the estimation space produces a $\hat{y}$ for which the sum of squared residuals has its least value? The squared distance from y^* to y is $\left(y-y^*\right)^T\left(y-y^*\right)=\left(y-X\hat{\beta}\right)^T\left(y-X\hat{\beta}\right)$. Thus, the least squares procedure applies when we choose the point in the estimation space that minimizes the squared distance. So, it is obvious that this request can be accomplished at the point $\hat{y}$ when we drop a perpendicular from y to the estimation space. The shortest distance $y-\hat{y}$ must be such that $X^T\left(y-\hat{y}\right)=X^T\left(y-X\hat{\beta}\right)=0$. Then, it would be obvious to show that the above implies that $\hat{\beta}$ is solved by $X^T y = X^T X\hat{\beta}$ which represents the ordinary least squares estimation.

5.2 Geometry of Ridge Regression

Figure 2 Estimation picture for Ridge regression



For the case of p=2, Figure 2 provides the picture of ridge regression. The constraint produces a feasible area (gray area). There are no corners for the contours to hit and hence zero solutions will rarely result. So, when p is large, the number of predictors would be very large and accordingly hard to interpret the model.

5.3 Geometry of Lasso

The criterion $\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij} \right)^2$ equals the quadratic function

$$\left\|Y - X\beta\right\|_2^2$$

$$= \left\| XX^{-1}\left(X^T\right)^{-1} X^T Y - X\beta \right\|_2^2$$

$$= \left\| X\left(X^T X\right)^{-1} X^T Y - X\beta \right\|_2^2$$

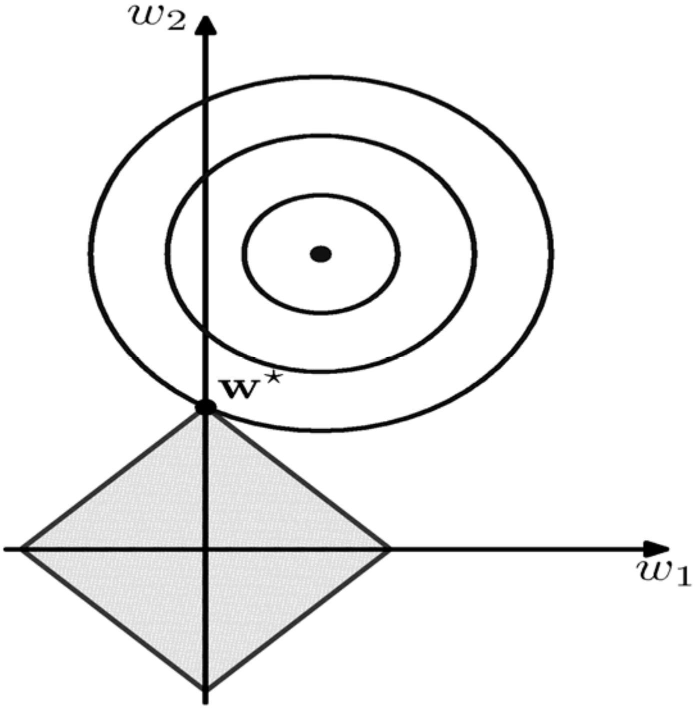
$$= \left\| X \overset{\wedge}{\beta}^{0} - X\beta \right\|_2^2$$

$$= \left\| X\left(\overset{\wedge}{\beta}^{0} - \beta \right) \right\|_2^2$$

$$= \left(\overset{\wedge}{\beta}^{0} - \beta \right)^T X^T X \left(\overset{\wedge}{\beta}^{0} - \beta \right)$$

The elliptical contours of this function can show why Lasso often produces coefficients that are exactly 0. When p=2, the constraint region is rotated square. The Lasso solution is the first place that the contours touch the square, and sometimes this will happens at a corner, which corresponds to a zero coefficient. Figure 3 shows the picture of Lasso when p=2.

Figure 3 Estimation picture for lasso



If we assume that the least squares estimates are both positive. (p=2) Then we can have the

Lasso estimates are $\hat{\beta}_j = \left(\left| \hat{\beta}_j^0 \right| - \gamma \right)^+$ where $\gamma = \left(\frac{\sum_j \left| \hat{\beta}_j^0 \right| - t}{2} \right)^+$

Since $\hat{\beta}_1 + \hat{\beta}_2 = t$, then we can solve the functions and get the solutions as following:

$$\hat{\beta}_1 = \left(\frac{t}{2} + \frac{\hat{\beta}_1 - \hat{\beta}_2}{2} \right)^+,$$

$$\hat{\beta}_2 = \left(\frac{t}{2} - \frac{\hat{\beta}_1 - \hat{\beta}_2}{2} \right)^+$$

Chapter 6

Standard Error of lasso

An approximate closed form estimate may be derived by writing the penalty $\sum_j |\beta_j|$ as $\sum_j \beta_j^{\,2} / |\beta_j|$. Hence, at the Lasso estimate, we may approximate the solution by a ridge regression of form $\beta^* = (X^T X + \lambda W^-)^{-1} X^T Y$ where W is diagonal matrix with diagonal elements $\left|\tilde{\beta}_j\right|$, W^- denotes the generalized inverse of W and λ is chosen that $\sum_j |\beta_j^*| = t$. The covariance matrix of the estimates could be approximated by

$$Cov(\beta^*) = \left(\left(X^T X + \lambda W^-\right)^{-1} X^T\right) Cov\left(Y\right) \left(\left(X^T X + \lambda W^-\right)^{-1} X^T\right)^T$$

$$= \left(X^T X + \lambda W^-\right)^{-1} X^T \overset{\wedge}{\sigma}^2 X \left(\left(X^T X + \lambda W^-\right)^T\right)^{-1}$$

$$= \left(X^T X + \lambda W^-\right)^{-1} X^T X \left(\left(X^T X + \lambda W^-\right)\right)^{-1} \overset{\wedge}{\sigma}^2$$

where $\overset{\wedge}{\sigma}^2$ is the estimate of the error variance. [1]

Chapter 7

Example—Hospital manpower data

The hospital manpower data come from Procedures and Analysis for Staffing Standards Development: Data/Regression Analysis Handbook. They are taken from seventeen Naval hospitals at various sites around the world. The goal is to produce an empirical equation that will estimate manpower needs for Naval hospitals. Former work was done by Raymond H. Myers by least squares method. I will use lasso method to do it.

A brief description of the predictor variables and response variable are as follows:

expos: Monthly X-ray exposures

days: Monthly occupied bed days

pop: Eligible population in the area divided by 1000

length: Average length of patients' stay in days

Table 1 Hospital manpower data

load	days	pop	length	hours
15.57	472.92	18	4.45	566.52
44.02	1339.75	9.5	6.92	696.82
20.42	620.25	12.8	4.28	1033.15
18.74	568.33	36.7	3.9	1603.62
49.2	1497.6	35.7	5.5	1611.37
44.92	1365.83	24	4.6	1613.27
55.48	1687	43.3	5.62	1854.17
59.28	1639.92	46.7	5.15	2160.55
94.39	2872.33	78.7	6.18	2305.58
128.02	3655.08	180.5	6.15	3503.93
96	2912	60.9	5.88	3571.89
131.42	3921	103.7	4.88	3741.4
127.21	3865.67	126.8	5.5	4026.52
252.9	7684.1	157.7	7	10343.81
409.2	12446.3	169.4	10.78	11732.17
463.7	14098.4	331.4	7.05	15414.94
510.22	15524	371.6	6.35	18854.45

Table 2 gives the least squares method results. The R code is attached. Of course, from the p-values, we can see not all predictors are significant. We can use the methods stated before to do the model selection.

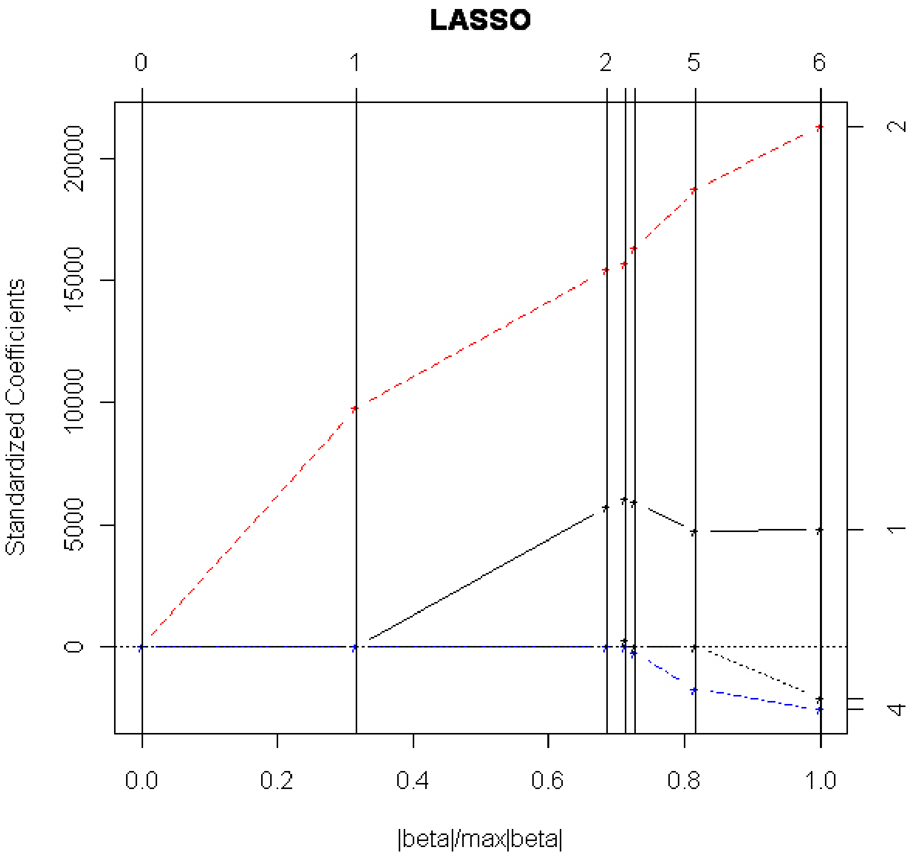
Table 2 Least squares estimate R output

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2032.17803	942.06971	2.157	0.0520
expos	0.05608	0.02036	2.755	0.0175 *
days	1.08837	0.15340	7.095	1.26e-05
pop	-5.00417	5.08070	-0.985	0.3441
length	-410.08088	178.07710	-2.303	0.0400 *

We can also use Lasso method to analyze this data. Figure 4 shows the lasso estimates.

Notice that the absolute value of standardized coefficient tends to 0 as the constraint goes to 0. In this example, we can also see that not all the curves decrease in a monotone fashion to 0. This lack of monotonicity is shared by ridge regression and subset regression. [1] We can use generalized cross-validation to choose the standardized bound $s = t / \sum \left| \hat{\beta}_j^{LSE} \right|$. For example, if $\hat{s} = 0.8$ is selected by generalized cross validation, then we can get the lasso estimate of predictors in table 3.

Figure 4 Lasso shrinkage of coefficients in the hospital manpower data



The lasso gave non-zero coefficients to expos, days, and length.

Table 3 Lasso estimates for predictors when $\hat{s} = 0.8$

Coefficient	expos	days	pop	length
estimate	0.05812114	0.93201364	0.00000000	-232.87973958

Chapter 8

Predication error and estimation of t

In this section we talk about two methods of choosing lasso parameter t: cross-validation and generalized cross-validation, and give some simple proof of the methods. These two methods are applicable in the 'X-random' case, where it is assumed that the observations (X, Y) are drawn from some unknown distribution. Yet, in real problems, we might choose these two methods in X-fixed case for convenience. [1]

8.1 Cross-validation method

Suppose that

$$Y = \eta(X) + \varepsilon$$

where $E(\varepsilon) = 0$ and $\operatorname{var}(\varepsilon) = \sigma^2$. The mean-squared error of an estimate $\hat{\eta}(X)$ is defined by

$$ME = E\left\{\hat{\eta}(X) - \eta(X)\right\},$$

the expected value taken over the joint distribution of X and Y, with $\hat{\eta}(X)$ fixed. A similar measure is the prediction error of $\hat{\eta}(X)$ given by

$$PE = E\left\{Y - \hat{\eta}(X)\right\}^2 = ME + \sigma^2.$$

Fivefold cross-validation is used in estimating the prediction error for lasso procedure. The lasso is indexed in terms of the normalized parameter $s = t / \sum\left|\hat{\beta}_j^{LSE}\right|$, and the prediction error is estimated over a grid of values of s from 0 to 1. The value yield the lowest estimated PE is selected. [1]

For linear model $\eta(X) = X\hat{\beta}$, the mean-squared error has the simple form

$$ME = \left(\hat{\beta} - \beta\right)^T V \left(\hat{\beta} - \beta\right)$$

Proof:

$$ME = E\left\{\hat{\eta}(X) - \eta(X)\right\}^2$$

$$= E\left\{X\hat{\beta} - X\beta\right\}^2$$

$$= E\left\{X\left(\hat{\beta} - \beta\right)\right\}^2$$

$$= Cov\left(X\left(\hat{\beta} - \beta\right)\right) + \left(E\left\{X\left(\hat{\beta} - \beta\right)\right\}\right)^2$$

$$= \left(\hat{\beta} - \beta\right)^T V \left(\hat{\beta} - \beta\right) + 0$$

$$= \left(\hat{\beta} - \beta\right)^T V \left(\hat{\beta} - \beta\right)$$

where V is the population covariance matrix of X.

8.2 Generalized cross-validation method

We write the constraint $\sum\left|\beta_j\right| \le t$ as $\sum\beta_j^2 / \left|\beta_j\right| \le t$. Then, the constraint is equivalent to adding a Lagrangian penalty $\lambda\sum\beta_j^2 / \left|\beta_j\right|$ to the residual sum of squares, with λ depending on t. Then, we can write the constrained solution $\tilde{\beta}$ as the ridge regression estimator

$$\tilde{\beta} = \left(X^T X + \lambda W^-\right)^{-1} X^T y$$

where $W = diag\left(\left|\tilde{\beta}_j\right|\right)$ and W^- is a generalized inverse. The number of effective parameters in the constrained fit $\tilde{\beta}$ may be approximate by [1]

$$p(t) = tr\left\{X\left(X^T X + \lambda W^-\right)^{-1} X^T\right\}.$$

Let rss(t) be the residual sum of squares for the constrained fit with constraint t. The generalized cross-validation style statistic

$$GCV(t) = \frac{1}{N}\frac{rss(t)}{\{1 - p(t)/N\}^2}.$$

Chapter 9

Summary and Conclusion

In this paper, firstly we have a review of several linear regression methods such as least squares method, ridge regression method, best subset method and the nowadays very popular lasso method. Then, two theorems, Lagrange multipliers and KKT conditions, are introduced. Thirdly, some proofs of lasso also introduced. We can get an intense understanding from the geometry of the methods. In the later parts, an example is also given. Last but not the least, the methods of choosing the constraint are stated.

From the analysis of methods and the example, we can see the advantage of lasso method. It can provide very good interpretation of predictors in real problem, especially in problems where there is a large number of predictors. What's more, the overall prediction accuracy is improved by sacrifice a little bias to reduce the variance of the predicted values.

Some new methods in lasso area are developed nowadays such as adaptive lasso, least angle regression algorithm and so on. This is a very interesting area and more work needs to be done. Maybe as the research work gets further, some more efficient algorithms could be invented.

Appendix

```
#R code

hop <- read.table("E:/project/hospital.txt",header=T,row.names=NULL,sep="  ")

attach(hop)

#least squares estimation

lm.hop<-lm(hours~expos+days+pop+length,data=hop)

summary(lm.hop)

x <- cbind(expos,days,pop,length)

#lasso method

object <- lars(x,hours,type="lasso")

plot(object)

fits <- predict.lars(object, x, type="fit")

coef <- predict(object, s=0.8, type="coef", mode="fraction")

coef
```

References

[1] Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B, Vol. 58, No.1, 1996.

[2] Matt Wand, Regression and Comparison, University of South Wales, 2005.

[3] Arthur E. Hoel and Robert W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, Vol. 42, No. 1, 2000.

[4] Gary C. McDonald, Ridge regression, John Wiley & Sons Inc, Volume 1, July/August2009.

[5] Raymond H. Myers, Classical and Modern Regression with Applications, Duxbury Press, 1990.

[6] http://wikipedia.org