1. Methods between Ridge and Lasso.

Consider the elastic-net optimization problem

$$\min_{\beta} \|y - X\beta\|^2 + \lambda(\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1)$$

Show how one can turn this into a lasso problem, using an augmented version of X and y.

Let us augment X with $I_{p\times p} * \theta$ to get $\tilde{X} = \begin{bmatrix} X \\ \theta I \end{bmatrix}$

Then $\tilde{X}\beta = \begin{bmatrix} X\beta \\ \theta\beta \end{bmatrix}$

Similarly, lets augment y with p zero values to get $\tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}$

$$\Rightarrow \|\tilde{y} - \tilde{X}\beta\|_2^2 = \left\| \begin{bmatrix} y - X\beta \\ \theta\beta \end{bmatrix} \right\|_2^2 = \|y - X\beta\|_2^2 + \theta^2\|\beta\|_2^2 \quad (1)$$

Following this augmentation, we estimate β for lasso by

$$\hat{\beta} = \arg\min_{\beta} \{\|\tilde{y} - \tilde{X}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1\} \quad (2)$$

Combining the two equations, we arrive at

$$\hat{\beta} = \arg\min_{\beta} \{\|y - X\beta\|_2^2 + \theta^2\|\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1\}$$

We transform the above by applying the following transformations:

- $\theta^2 = \lambda\psi$

  Then $\theta = \sqrt{\lambda\psi}$
- $\tilde{\lambda} = \lambda(1-\psi)$

To optimize the problem, let $\tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}$, $\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda\psi} I \end{bmatrix}$ and $\tilde{\lambda} = \lambda(1-\psi)$

The lasso problem is solved as such ■

2. Closed forms of Lasso, Ridge, and Best Subset Selection.

Consider a special case of linear regression

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

where the columns of X are orthonormal, that is $X^TX = I$. Let $\hat{\beta}$ denote the least square estimator and $\hat{\beta}_{(M)}$ denote the M^{th} largest absolute value of $\hat{\beta}$, that is $|\hat{\beta}_{(1)}| \geq |\hat{\beta}_{(2)}| \geq \ldots \geq |\hat{\beta}_{(p)}|$. Show the corresponding formulas for Best Subset with size M, Ridge, and Lasso estimator

Estimator	Formula
Best Subset (Size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j / (1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Given that X are orthonormal, $X^TX = I$

For Least Squares, $\hat{\beta} = (X^TX)^{-1}X^Ty = X^Ty$

(i) Best Subset Selection

First we minimize the residual sum of squares:

$$\Rightarrow (y - X\hat{\beta})^T(y - X\hat{\beta}) = y^Ty - 2\hat{\beta}X^Ty + \hat{\beta}^2$$

Substitute $\hat{\beta} = X^Ty$

$$\Rightarrow y^Ty - 2(X^Ty)^2 + (X^Ty)^2 =$$

$$= y^Ty - (X^Ty)^2$$

$$= y^Ty - \|\hat{\beta}\|^2$$

Thus, the minimal residuals occurs when $\|\hat{\beta}\|^2$ is maximized. The best subset with M predictors adheres to the following property

$$\hat{\beta}_j = I(|\hat{\beta}_j| \geq |\beta_{(M)}|).$$

In real terms this equates to choosing M largest coefficients by their absolute value ■

(ii) Ridge

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \{(y - X\beta)^T(y - X\beta) + \lambda \beta^T \beta\}$$

Distributing:

$$= \underset{\beta}{\operatorname{argmin}} \{y^T y - 2\beta X^T y + X^T X \beta^T \beta + \lambda \beta^T \beta\}$$

$$= \underset{\beta}{\operatorname{argmin}} \{y^T y - 2\beta X^T y + \beta^T \beta (I + \lambda)\}$$

Differentiate w.r.t β and set to 0:

$$-2X^T y + 2\beta(I + \lambda) = 0$$

Substitute OLS definition $\hat{\beta}^{ols} = X^T y$

$$-2\hat{\beta}^{ols} + 2\beta(I + \lambda) = 0$$

$$\beta = \frac{\hat{\beta}^{ols}}{I + \lambda} \Longrightarrow \beta_i = \frac{\hat{\beta}_i^{ols}}{1 + \lambda}$$ ■

(iii). Closed form of Lasso

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

$$= \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \left[y^T y - 2\beta X^T y + X^T X \beta^T \beta \right] + \lambda \|\beta\|_1 \right\}$$

$X^T X = I$

$$= \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \left[y^T y - 2\beta \hat{\beta}^{ols} + \beta^T \beta \right] + \lambda \|\beta\|_1 \right\}$$

Take the derivative w.r.t β and set to 0:

$$-\frac{2}{2}\hat{\beta}^{ols} + \frac{2}{2}\beta_i + \lambda \beta_i = 0$$

We must examine several cases

Case (1) $\hat{\beta}_j^{ols} \geq 0$ Then $\beta_j \geq 0$

Our equation changes to

$$-\hat{\beta}_j^{ols} + \beta_j + \lambda = 0$$

$$\beta_j = \hat{\beta}_j^{ols} - \lambda$$

Note: Our assumption is that $\beta_j \geq 0$

Then $\beta_j = \text{sign}(\hat{\beta}_j)^{ols} (\hat{\beta}_j^{ols} - \lambda)_+$

Case ② $\beta_j < 0$

The expression changes to:

$$-\hat{\beta}_j^{OLS} + \beta_j - \lambda > 0$$

$$\Rightarrow \hat{\beta}_j^{Lasso} = \hat{\beta}_j^{OLS} + \lambda$$

We note that $\hat{\beta}_j^{OLS} + \lambda$ must be non-negative

Thus we derived the equation:

$$\hat{\beta}_j^{Lasso} = sign(\hat{\beta}_j^{OLS}) * (|\hat{\beta}_j^{OLS}| - \lambda)_+$$ ■