

ESL 3.13

Let z_m with principal comp

$$z_m = \sum_{j=1}^p x_j y_{jm}$$

$$\begin{aligned} \hat{y} &= \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j z_j \\ &= \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j \sum_{m=1}^n x_{jm} v_m \\ &= \hat{\theta}_0 + \sum_{m=1}^n \left(\hat{\theta}_1 x_{1m} + \dots + \hat{\theta}_p x_{pm} \right) v_m \\ &= \hat{\theta}_0 + \sum_{m=1}^n \hat{\beta}^{\text{OLS}}(x_m) v_m \\ &= \hat{\theta}_0 + \sum_{m=1}^n \hat{\beta}^{\text{OLS}}(x_m) v_m \end{aligned}$$

To show that $\hat{\beta}^{\text{OLS}}(p) = \hat{\beta}^{\text{LS}}$:
Consider the singular value decomposition of X :

$$X = U D V^T \quad D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p & & \\ & & & 0 & \dots \end{bmatrix}$$

$$V^T = [v_1 \dots v_p]$$

- both U and V orthogonal

$$V^T V = I_p \quad U^T U = I_n$$

- D is diagonal, $D^T = D$

$$\begin{aligned} X^T X &= (V D U^T)(U D V^T) \\ &= V D^2 V^T \end{aligned}$$

The least square coefficients can then be written as:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (V D^2 V^T)^{-1} (U D V^T)^T y \\ &= (D^2 V V^T)^{-1} V D U^T y \\ &= D^{-2} V D U^T y \\ &= V D^{-1} U^T y \end{aligned}$$

Now, we show that $\hat{\beta}^{\text{OLS}} = \hat{\beta}$:

$$\begin{aligned} Z &= [z_1 \dots z_p] \\ &= [x v_1 \dots x v_p] \\ &= X [v_1 \dots v_p]^T \\ &= X V \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \\ &\quad \begin{bmatrix} x_j & x_j \end{bmatrix} \begin{bmatrix} v_{j1} \\ v_{j2} \end{bmatrix} \\ &\quad n \times p \end{aligned}$$

$$\begin{aligned} Z &= (U D V^T) V \\ &= U D \end{aligned}$$

$$\begin{aligned} \hat{\theta} &= (Z^T Z)^{-1} Z^T y \\ &= (D U^T U D)^{-1} (D U^T) y \\ &= D^{-2} D U^T y \\ \hat{\theta} &= D^{-1} U^T y \end{aligned}$$

From part 1,

$$\begin{aligned} \hat{\beta}^{\text{OLS}}(p) &= \sum_{m=1}^p \hat{\theta}_m v_m \\ &= \begin{bmatrix} v_1 & \dots & v_p \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{bmatrix} \\ &= V \hat{\theta} \\ &= V D^{-1} U^T y \\ &= \hat{\beta} \end{aligned}$$

ESL 3.16

All three methods solve a constrained minimization problem,

$$\hat{\beta}^* = \arg \min \{ \text{RSS}(\beta) \text{ s.t. } \beta \in \mathcal{B} \}$$

ESL 3.16

All three methods solve a constrained minimization problem,

$$\hat{\beta}^* = \arg\min_{\beta} \{ \text{RSS}(\beta) \text{ s.t. } \beta_j \leq s_j \}$$

If $X = [x_1 \dots x_p]$ all columns are orthogonal, then

$$x_i^T x_j = 0 \quad \text{if } i \neq j$$

$$x_i^T x_i = 1$$

This means that

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

$$= X^T y$$

$$\hat{\beta}_j = x_j^T y$$

which is the univariate regression solution.

(a) rewrite the minimization problem as follows:

$$\arg\min_{\beta} \text{RSS}(\beta) =$$

$$\arg\min_{\beta} (y - X\beta)^T (y - X\beta) =$$

$$\arg\min_{\beta} (y^T y + \underbrace{\beta^T X^T X \beta}_{= I \beta} - 2y^T X \beta) =$$

$$\arg\min_{\beta} (\beta^T \beta - 2\hat{\beta}^T \beta) + \underbrace{y^T y}_{\text{constant}}$$

$$\arg\min_{\beta} (\beta - \hat{\beta})^T (\beta - \hat{\beta}) = \sum_j (\beta_j - \hat{\beta}_j)^2$$

does not depend on β_j value

① Best subset (BS):

Since all columns in X are orthogonal, we know that

$$\hat{\beta}_j^{BS} = \begin{cases} \hat{\beta}_j & \text{if } x_j \text{ is included} \\ 0 & \text{if not} \end{cases}$$

$$\hat{\beta}^{BS} = \arg\min_{\beta} \left\{ \text{RSS}(\beta) \mid \sum_{j=1}^p \mathbb{1}(\beta_j > 0) \leq s \right\}$$

number of coefficients

$$= \arg\min_{\beta} \left\{ \sum_j (\beta_j - \hat{\beta}_j)^2 \mid \sum_{j=1}^p \mathbb{1}(\beta_j > 0) \leq s \right\}$$

Here, we will minimize $\sum_j (\beta_j - \hat{\beta}_j)^2$ by excluding from the set (set to zero) the smallest LS coefficients

↳ if we include x_j , then $(\beta_j - \hat{\beta}_j)^2 = 0$
if we exclude x_j , then $(\beta_j - \hat{\beta}_j)^2 = \hat{\beta}_j^2$

(it's expected coefficient)
 coefficient

This we set

$$\hat{\beta}_j^{BS} = \begin{cases} \hat{\beta}_j & \text{if } |\hat{\beta}_j| > \hat{\beta}_{(s+1)} \\ 0 & \text{otherwise} \end{cases}$$

Ridge

$$\hat{\beta}^{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

$$= (I + \lambda)^{-1} \hat{\beta}$$

$$= \frac{1}{(1+\lambda)} \hat{\beta}$$

Lasso:

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \left\{ \sum_j (\beta_j - \hat{\beta}_j)^2 + \lambda \sum_j |\beta_j| \right\}$$

can solve it individually:

for each j :

$$\hat{\beta}_j^{Lasso} = \arg\min_{\beta_j} \left\{ (\beta_j - \hat{\beta}_j)^2 + \lambda |\beta_j| \right\}$$

Lasso:

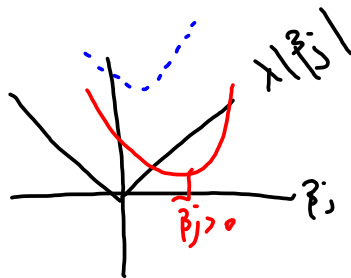
$$\hat{\beta}^{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_j (\tilde{y}_j - \beta_j)^2 + \lambda \sum_j |\beta_j| \right\}$$

can solve it individually:

For each j :

$$\hat{\beta}_j^{\text{Lasso}} = \underset{\beta_j}{\operatorname{argmin}} \underbrace{\left\{ (\tilde{y}_j - \beta_j)^2 + \lambda |\beta_j| \right\}}_{F(\beta)}$$

$$F'(\beta) = \begin{cases} 2(\tilde{y}_j - \beta_j) + \lambda & \text{if } \tilde{y}_j \geq 0 \\ 2(\tilde{y}_j - \beta_j) - \lambda & \text{if } \tilde{y}_j < 0 \end{cases}$$



if $\hat{\beta}_j \geq 0$:

$$\hat{\beta}_j^{\text{Lasso}} = \left\{ \hat{\beta}_j - \frac{\lambda}{2}, 0 \right\}$$

if $\hat{\beta}_j < 0$:

$$\hat{\beta}_j^{\text{Lasso}} = \min \left\{ \hat{\beta}_j + \frac{\lambda}{2}, 0 \right\}$$