6 Regularization

Recall the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

Note

- If $p \gg 0$ but $p < n$ then the OLS estimator

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

- If $p > n$ then the OLS estimates are not identifiable b/c

$$\underbrace{(X^T X)}_{\text{rank deficient} \Rightarrow \text{not invertible}} \beta = X^T y$$

So there are infinitely many solutions.

$\boxed{\text{Options}}$

- Do subset selection
- Include all predictors, but regularize their effects to shrink them toward zero.

$\boxed{\text{Motivation}}$ Given samples $(\underline{x}_1, y_1), \ldots, (\underline{x}_n, y_n)$

OLS minimizes

$$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$$

regularization adds a penalty $\Rightarrow$

$$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) + \lambda P(\underline{\beta})$$

where $P(\underline{\beta})$ is a penalty / regularization term that grows with the size of $\underline{\beta}$, and shrinks to zero when $\underline{\beta} = \underline{0}$

$\lambda \geq 0$ is a smoothing | shrinkage | complexity | regularization parameter.

$\boxed{\text{Note}}$ why does $P(\underline{\beta})$ control the size of the model?

- Case 1: $\underline{\beta} = \underline{0}$, so $P(\underline{\beta}) = 0 \Rightarrow Y$ does not depend on $\underline{x}$ at all, so we get a small model

- Case 2: $\underline{\beta}$ big in every feature, $P(\underline{\beta}) \gg 0$
$\Rightarrow Y$ depends strongly on all features.

6.1 Ridge Regression

β_0 only measures the avg value of Y, so should not be penalized.

If we use centered features $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$Y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i \qquad i = 1, \dots, n$$

What is $\hat{\beta}_0$?

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 (x_i - \bar{x}) \right)^2$$

$$= -2 \sum_{i=1}^{n} (Y_i - \beta_0) + 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) \quad 0 \checkmark$$

$$\sum (x_i - \bar{x})$$
$$= \sum x_i - \sum \bar{x}$$
$$= \sum x_i - n\bar{x}$$
$$= \sum x_i - \sum x_i = 0$$

$$= -2n\bar{Y} + 2n\beta_0$$

set 0

$$\Rightarrow \boxed{\bar{Y} = \hat{\beta}_0} \text{ is OLS.}$$

$\Rightarrow$ throughout the section we will center features + response

$\boxed{\text{Note}}$ The basic idea behind ridge regression is

$$P(\beta) = \|\beta\|_2^2 = \beta^T \beta = \sum_{j=1}^{p} \beta_j^2$$

where now $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$ $\longrightarrow$ no β_0 term b/c we'll use centered features.

$\boxed{\text{Problem}}$ $p=2$ x_1 = budget of movie in \$s ($\sim$ 1000000s ish)

x_2 = rating of movie ($\sim$ 1-10 ish)

$$\|\beta\|_2^2 = \beta_1^2 + \beta_2^2 \quad \text{units?}$$

Nonsensical $\Rightarrow$ need to remove the units of features.

$\boxed{\text{Warning / Note / Convention}}$

For the rest of the chapter, we will assume:

- $\hat{\beta}_0 = \bar{Y}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$ p features, no β_0

- observations are centered: $Y_i \to Y_i - \bar{Y}$

- Features have been centered & scaled:

$$x_i \to \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}}$$ [x is unitless]

Implications of assumptions

- Y is mean zero when $(x_1, \dots, x_p) = (0, \dots, 0)$
- x_is are unitless
- $$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2}} \right]^2$$

new xs

$\rightarrow$ old xs

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2} = n$$

DEF The ridge regression estimator for $\underline{\beta}$ minimizes

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

$$= (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) + \lambda \|\underline{\beta}\|_2^2$$

design matrix has no 1st column of 1s.

Ex $p=1$ model $Y = \beta_1 x + \varepsilon$