REGULARIZATION

1. Multiple linear regression

Convention: We view $\mathbb{R}^k$ as a subset of $\mathbb{R}^{k+1}$ via the following identification

$$v \in \mathbb{R}^k \quad \longleftrightarrow \quad (1, v) \in \mathbb{R}^{k+1}. \tag{1}$$

$p-1$ predictor variables:

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{1\times(p-1)} \times \mathbb{R}$$

Viewing x_i as an element of $\mathbb{R}^{1\times p}$ via (1), define:

$$x := \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n\times p}, \quad y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n\times 1}$$

For $\beta \in \mathbb{R}^{p\times 1}$, consider the equation:

$$x\beta = y$$

Equivalently:

$$\beta_0 + \beta_1 x_{i,1} + \cdots \beta_{p-1} x_{i,p-1} = y_i, \quad i = 1, \ldots, n.$$

Recall: The *column space of* x is the subspace $C(x)$ of $\mathbb{R}^{n\times 1}$ characterized by any of the following equivalent conditions:

- $C(x)$ is the set of all linear combinations of the columns of x
- $C(x) = \{x\beta : \beta \in \mathbb{R}^{p\times 1}\}$
- $C(x) = \{y \in \mathbb{R}^{n\times 1} : x\beta = y \text{ has a solution}\}$

$C(x)$ is also called the *image of* x.

Let $\widehat{y} \in \mathbb{R}^{n\times 1}$ be the vector characterized by any of the equivalent conditions:

- $\widehat{y} = \underset{z\in C(x)}{\operatorname{argmin}} \|z - y\|$
- $\widehat{y}$ is the vector in the column space of x closest to y.
- $\widehat{y}$ is the orthogonal projection of y onto the column space of x.

In partricular, $x\beta = \widehat{y}$ has a solution.

2. The case $\operatorname{rank}(x) = p$

Suppose $\operatorname{rank}(x) = p$. Then $\beta \mapsto x\beta$ maps $\mathbb{R}^{p\times 1}$ bijectively onto $C(x)$ and, therefore, there is a unique vector $\widehat{\beta} \in \mathbb{R}^{p\times 1}$ — the *least squares solution of* $x\beta = y$ — such that

$$x\widehat{\beta} = \widehat{y}.$$

The vector $\widehat{\beta}$ is characterized by the fact that it minimizes the sum of squared errors in approximating y by a vector of the form $x\beta$:

$$\widehat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^{p\times 1}} \|x\beta - y\|^2$$

Since $\operatorname{rank}(x) = p$, the matrix $x^T x \in \mathbb{R}^{p\times p}$ is invertible and the system

$$x^T x\beta = x^T y$$

has unique solution; this solution is just $\widehat{\beta}$:

$$\widehat{\beta} = (x^T x)^{-1} x^T y$$

Thus,

$$\widehat{y} = x\widehat{\beta} = Py,$$

where

$$P := x(x^T x)^{-1} x^T.$$

The matrix P is called the *projection matrix* because it describes orthogonal projection from $\mathbb{R}^{n\times 1}$ onto $C(x)$.

If we view the y_i as realizations of random variable Y_i and let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n, \end{bmatrix}$$

then we may view

$$\widehat{\beta} = \widehat{\beta}(Y_1, \dots, Y_n) = (x^T x)^{-1} x^T Y$$

as an estimator.

Theorem 1. *Suppose* $\operatorname{rank}(x) = p$ *and*

$$Y \sim N(x\beta, \Sigma).$$

Then $\widehat{\beta}$ *is an unbiased estimator of* β.

Proof. Use the linearity of expectation:

$$\begin{aligned}\mathbb{E}\,\widehat{\beta} = \mathbb{E}\left[(x^T x)^{-1} x^T Y\right] &= (x^T x)^{-1} x^T\,\mathbb{E}\,Y \\ &= (x^T x)^{-1} x^T (x\beta) = (x^T x)^{-1}(x^T x)\beta = I\beta = \beta\end{aligned}$$

□

$$
\begin{aligned}
\operatorname{Var}\widehat{\beta} &= \operatorname{Var}(x^T x)^{-1}x^T Y \\
&= (x^T x)^{-1}x^T(\operatorname{Var}Y)((x^T x)^{-1}x^T)^T \\
&= (x^T x)^{-1}x^T(\sigma^2 I)x(x^T x)^{-1} \\
&= \sigma^2(x^T x)^{-1}(x^T x)(x^T x)^{-1} \\
&= \sigma^2(x^T x)^{-1}
\end{aligned}
$$

3. The case $\operatorname{rank}(x) \leq p$

We consider a *regularized* version of multiple linear regression. Let $\lambda > 0$ and consider the problem of minimizing

$$\mathrm{SSE}_\lambda(\beta) := \|x\beta - y\|^2 + \lambda^2\|\beta\|^2$$

Let

$$\xi \;:= \begin{bmatrix} x \\ \lambda I^{p\times p} \end{bmatrix} \in \mathbb{R}^{(n+p)\times p}, \quad \eta := \begin{bmatrix} y \\ 0^{p\times 1} \end{bmatrix} \in \mathbb{R}^{(n+p)\times 1}$$

and consider the equation

$$\xi\beta = \eta.$$

The columns of ξ are linearly independent (why?), so $\operatorname{rank}(\xi) = p$. Therefore, by the discussion of the previous section, $\xi^T\xi$ is invertible and

$$\widehat{\beta}_\lambda := (\xi^T\xi)^{-1}\xi^T\eta$$

minimizes

$$\|\xi\beta - \eta\|^2 = \left\| \begin{bmatrix} x\beta - y \\ \lambda\beta \end{bmatrix} \right\|^2 = \|x\beta - y\|^2 + \lambda^2\|\beta\|^2 = \mathrm{SSE}_\lambda(\beta).$$

We have:

$$
\begin{aligned}
\xi^T\xi &= \begin{bmatrix} x^T & \lambda I \end{bmatrix} \begin{bmatrix} x \\ \lambda I \end{bmatrix} \\
&= x^T x + \lambda^2 I, \\
\xi^T\eta &= \begin{bmatrix} x^T & \lambda I \end{bmatrix} \begin{bmatrix} y \\ 0^{p\times 1} \end{bmatrix} \\
&= x^T y
\end{aligned}
$$

Therefore,

$$\widehat{\beta}_\lambda = (x^T x + \lambda^2 I)^{-1}x^T y.$$

Let

$$W_\lambda = (x^T x + \lambda^2 I)^{-1}x^T x.$$

Theorem 2. *Suppose* $\operatorname{rank}(x) = p$*, so that* $\widehat{\beta}$ *is defined. Then*

$$\beta_\lambda = W_\lambda\widehat{\beta}.$$

Proof. Just compute.

$$\begin{aligned}
W_\lambda \widehat{\beta} &= (x^T x + \lambda^2 I)^{-1} x^T x \widehat{\beta} \\
&= (x^T x + \lambda^2 I)^{-1} x^T x (x^T x)^{-1} x^T y \\
&= (x^T x + \lambda^2 I)^{-1} x^T y \\
&= \widehat{\beta}_\lambda.
\end{aligned}$$

□