Lecturer: Prof. Homrighausen Scribe: Ahmed Alawami

Let $\mathbb{X} \in \mathbb{R}^{n \times p}$. There are three regimes of interest:

1. **Classical**: $n \gg p$ and n small.
2. **Big Data**: $n \gg p$ and n large.
3. **High Dimensional**: $n \leq p$.

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \hat{R}(f_\beta) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 \tag{1}$$

Note. There is a unique solution only if $\text{rank}(\mathbf{X}) = p$

It follows that:

$$\hat{f}(\mathbf{X}) = \mathbf{X}^T \hat{\boldsymbol{\beta}}_{LS} = \mathbf{X}^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y} \tag{2}$$

Note. In general, $\hat{\boldsymbol{\beta}}_{LS} = \mathbb{X}\dagger$ where $\mathbb{X}\dagger$ is the Moore-Penrose pseudo inverse.

and the fitted values are: $\mathbb{X}\hat{\beta}_{LS} = \mathbf{HY}$, where $\mathbf{H}$ is the orthogonal projection onto the column space of $\mathbb{X}$.

Note the following:

$$\begin{aligned} \mathbb{E}\hat{\boldsymbol{\beta}}_{LS} &= \boldsymbol{\beta} \\ \mathbb{V}\hat{\boldsymbol{\beta}}_{LS} &= \mathbb{X}\dagger \, \mathbb{V}\mathbf{Y} (\mathbb{X}\dagger)^T \end{aligned}$$

Note. We are estimating β with respect to the space:

$$\{f : \text{ There exists } \beta \text{ where } f(\mathbf{X}) = \beta^T \mathbf{X}\} \tag{3}$$

1 Gauss-Markov Theorem

Theorem 1.1. *If $\mathbb{E}\mathbf{y} = \mathbf{X}\beta$ and $cov(\mathbf{y}) = \sigma^2\mathbf{I}$, the least-squares estimators $\hat{\beta}_j, j = 1, \cdots, k$ have minimum variance among all linear unbiased estimators.*

Note, it is also the maximum likelihood estimator, if we assume equal variances and normality of error.

But fitting least-squares model, can lead to poor prediction and estimation performance.

1.1 Example of LS problems

Write $\mathbb{X}$ using SVD as $\mathbb{X} = \mathbf{UDV}^T$. So,

$$\mathbb{V}\hat{\boldsymbol{\beta}}_{LS} \propto \left(\mathbb{X}^T\mathbb{X}\right)^{-1} = \mathbf{VD}^{-2}\mathbf{V}^T$$

Therefore,

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_2^2 = \text{trace}(\mathbb{V}\hat{\boldsymbol{\beta}}) \propto \sum_{j=1}^{p} \frac{1}{d_j^2}$$

Note, we can get arbitrarly bad behavior if $d_p \approx 0$.

1.2 Vandermonde Matrix

The *Vandermonde Matrix* arises in polynomial regression. An order p Vandermonde matrix as follows:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{p-1} \end{bmatrix}$$

One appealing feature of the Vandermonde Matrix is it has the following simple determinant form:

$$\det(V) = \prod_{1 \le i \le j \le p} (x_j - x_i)$$

1.3 Spectral (Operator) Norm

The *spectral norm* is a way to measure the 'size' of a linear operator.

Definition 1.2. *Let* $\mathbf{A} : \mathbf{V} \to \mathbf{W}$ *be a continuous linear map. The* ***spectral norm****,* $\|\cdot\|_{spectral}$*, is defined as follows:*

$$\|\mathbf{A}\|_{spectral} = inf\{c \ge 0 | \|\mathbf{Av}\| \le c\|\mathbf{v}\| \text{ for all } \mathbf{v} \in \mathbf{V}\}$$

2 Big Data Regime

The computational complexity scales extremely quickly. For example, the least squares estimator scales linearly with data and quadratically with the parameters. This means that some classical procedures are no longer feasible for large data sets.

3 High Dimensional Regime

In addition to the computational problems of big data, there is a *rank problem*.

3.1 Rank Problem

Suppose $\mathbb{X} \in \mathbb{R}^{\aleph \times \iota}$ and $p > n$. Therefore, $\text{rank}(\mathbb{X}) = n$ and $\mathbb{X}\hat{\boldsymbol{\beta}}$. That is, the equation can be solved exactly and moreoever it has an infinite number of solutions.