STAT 499: Undergraduate Research

Week 1: Introduction of the linear model

Facheng Yu Compiled on: 2024-01-25, 16:38:20

1.1 Least squares estimator

Suppose that $x \in \mathbb{R}^d$ is a random vector and $y \in \mathbb{R}$ is a random variable.

Mean squared error. For any estimator $f(x)$ for y, the mean squared error is defined as

$$MSE_f = E[(f(x) - y)^2].$$

For a given model class $\mathcal{F}$, the least squares estimator is defined as

$$f^* = argmin_{f \in \mathcal{F}} MSE_f = argmin_{f \in \mathcal{F}} E[(f(x) - y)^2].$$

By definition, it is easy to see that least squares estimator is also the smallest variance estimator.

Empirical MSE. When the distribution of x and y is unknown, we cannot compute MSE directly. Instead, we can compute the empirical mean squared error using observed samples $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\widehat{MSE}_f = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2.$$

For a given model class $\mathcal{F}$, the empirical least squares estimator is defined as

$$\hat{f} = argmin_{f \in \mathcal{F}} \widehat{MSE}_f = argmin_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2.$$

Examples: Linear model. Suppose that the model class $\mathcal{F}$ is defined as

$$\mathcal{F} = \{f : f(x) = \alpha + x^T \beta, \alpha \in \mathbb{R}, \beta \in \mathbb{R}^d\}.$$

(i): The mean squared error is a function of α and β:

$$MSE_f = E[(\alpha + x^T \beta - y)^2] =: R(\alpha, \beta).$$

Let $z = (1, x^T)^T$, and $\gamma = (\alpha, \beta)$. The least squares estimator $f^*(x) = \alpha^* + x^T \beta^* = z^T \gamma^*$ is given by

$$\gamma^* = argmin_{\gamma \in \mathbb{R}^{d+1}} E[(z^T \gamma - y)^2].$$

By differentiating, we would have

$$(\alpha^*, \beta^*) = \gamma^* = E[zz^T]^{-1} E[zy].$$

(ii): The empirical least squares estimator $\hat{f} = \hat{\alpha} + x^T \hat{\beta} = z^T \hat{\gamma}$ is given by

$$\hat{\gamma} = argmin_{\gamma \in \mathbb{R}^{d+1}} \frac{1}{n} \sum_{i=1}^n (z_i^T \gamma - y_i)^2.$$

Let $Z = (z_1, \ldots, z_n)^T$ and $Y = (y_1, \ldots, y_n)^T$, then we have

$$\hat{\gamma} = argmin_{\gamma \in \mathbb{R}^{d+1}} \frac{1}{n} \|Z\gamma - Y\|_2^2.$$

By differentiating, we would have

$$(\hat{\alpha}, \hat{\beta}) = \hat{\gamma} = (Z^T Z)^{-1}(Z^T Y).$$

So, one natural question is that when would $E[zz^T]$ or $Z^T Z$ be invertiable?

1.2 Invertible $E[zz^T]$ and ZZ^T

Invertible $E[zz^T]$. We first consider when $E[zz^T]$ is invertiable.

Recap that $z = (1, x^T)^T$. Assume that $x \sim N_d(0, \Sigma)$, then we have

$$E[zz^T] = Cov(z, z) + E[z]E[z^T] = \begin{pmatrix} 1 & 0 \\ 0 & \Sigma \end{pmatrix}.$$

Since $\mathsf{det}(E[zz^T]) = \mathsf{det}(\Sigma)$, $E[zz^T]$ would be invertible if Σ is invertible. Intuitively, the weaker correlations between each pair of covariates is, the more likely $E[zz^T]$ being invertiable would be.

Invertible ZZ^T. Now we consider when $Z^T Z$ is invertiable.

(i) $n >> d$. Consider the first $d+1$ observations $z_1, \ldots, z_{d+1}$. If $M = \mathsf{span}\{z_1, \ldots, z_{d+1}\} \neq R^{d+1}$, then with high probability, the next observation z_{d+2} would have non-zero projection on the orthogonal space M, which contributes to additional ranks. So when $n >> d$, $Z^T Z$ is invertible with high probability.

(ii) $d >> n$. Since $\mathsf{rank}(Z^T Z) \leq n << d$, $Z^T Z \in \mathbb{R}^{1+d \times 1+d}$ cannot be invertible so we cannot simply use $(Z^T Z)^{-1}(Z^T Y)$ as the solution.

In case (ii), such a question is considered:

$$Z^T Z\gamma = Z^T Y.$$

By linear algebra we know that, the system above has either no solution or infinitely many solutions. To get a unique solution, one can "throw" some features and consider a reduced question with $d' << n$, then it comes back to case (i).

1.3 Sparsity models

Now we continuing to consider case (ii) where $d > n$, and take the same notations as in the Wainwright's book. Let $\theta^* \in \mathbb{R}^d$ be an unknown regression vector. Suppose that we observe $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d}$ via the linear model:

$$y = X\theta^* + w$$

Hard sparsity The support set of θ^* is defined as

$$S(\theta^*) := \{j \in \{1, \ldots, n\} : \theta_j^* \neq 0\}.$$

The hard sparsity requires $s := |S(\theta^*)|$ substantially smaller than d. Under the sparsity assumption, we may have a unique linear solution of the least squares estimator.

Example 7.3 Consider polynomial functions in a scalar variable $t \in \mathbb{R}$ of degree k, say of the form

$$f_\theta(t) = \theta_1 + \theta_2 t + \cdots + \theta_{k+1} t^k$$

Suppose that we observe n samples $\{(y_i, t_i)\}_{i=1}^n$ via $y_i = f_\theta\left(t_i\right) + w_i$. Define the $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^k \end{bmatrix}$$

and we have $y = \mathbf{X}\theta + w$.