Linear Regression Model

2021.09.12

1. Linear Regression model

$$Matrix\ form : Y = X\beta + \epsilon$$
$$Scalar\ form\ \ : y_i = f(\mathbf{x}_i) + \epsilon_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$
$$\epsilon_i \sim iid\ \ N(0, \sigma^2)$$

- $X = (X_1, \cdots, X_p)$: Input, Predictor variables
- Y : Output, Response variable
- β : Unknown parameter
- ϵ : error term.

2. Least Square estimate of β

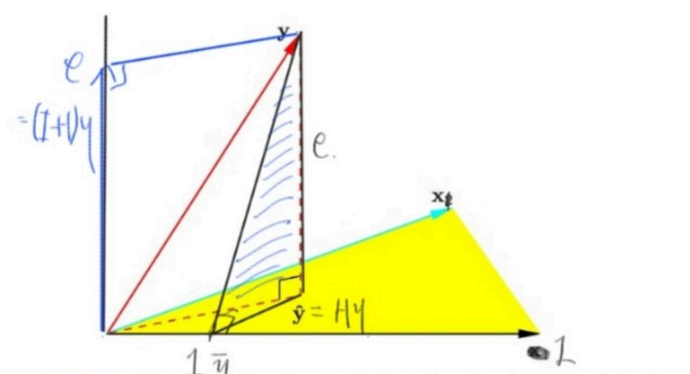
- Minimize the Residual sum of squares (RSS)

$$\underset{\beta}{argmin}\ RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j)^2$$
$$= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$$

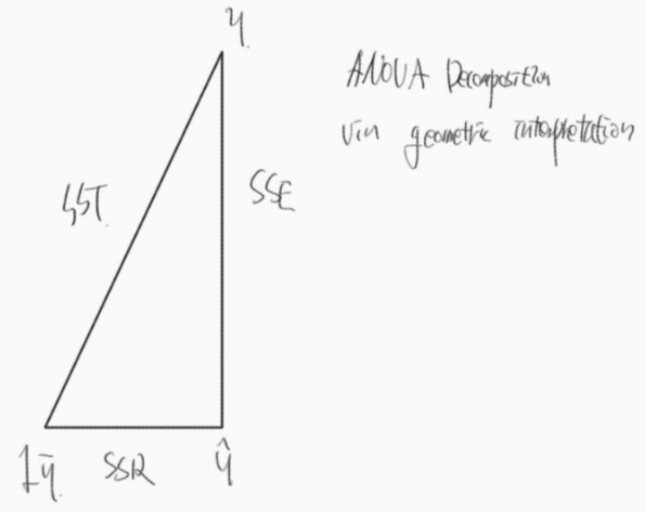
- estimation & prediction when X is non-singular

> $\hat{\beta}^{LSE} = (X^TX)^{-1}X^T\mathbf{y}\ \ \sim\ \ N(\beta, (X^TX)^{-1}\sigma^2)$ -> 식으로 유도
>
> $\hat{y} = X\hat{\beta}^{LSE} = X(X^TX)^{-1}X^T\mathbf{y} = \mathbf{H}\mathbf{y}$
>
> $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
>
> $\frac{(N-p-1)\hat{\sigma^2}}{\sigma^2}\ \sim \chi^2_{N-p-1}\ , where\ \ \hat{\sigma^2} = \sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{(N-p-1)}(= s^2)$ -> 식으로 유도
>
> $\hat{\beta}, \sigma^2$ are independent to each other

3. Further discussion about Linear regression



IGURE 3.2. *The N-dimensional geometry of least squares regression with two edictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane anned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$. *The projection* $\hat{\mathbf{y}}$ *represents the vector the least squares predictions*



- Column space and Null space

> Column space of A is a space spanned by columns of A
>
> Null space of A is a space spanned by $\mathbf{x}$ which satisfies $A\mathbf{x} = \mathbf{0}$
>
> *C(A)* $\perp$ *N(A)*

- Hat matrix

> $H = X(X^TX)^{-1}X^T$
>
> Hat matrix is projection marix to column space of X
>
> H is symmetric($H^T = H$) and idempotent($H^2 = H$)
>
> $(I - H)$ is projection matrix to Null space of X , and this is also symmetric and idempotent

- ANOVA decomposition of regression model

> $SST = \sum_{i=1}^{N}(y_i - \bar{y})$, $df = N$
>
> $SSR = \sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)$, $df = p + 1$
>
> $SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)$, $df = N - p - 1$
>
> $SST^2 = SSR^2 + SSE^2$

- Gauss Markov Theorem

> $Var(\hat{\theta}^{LSE}) \leq Var(\widetilde{\theta})$ -> 유도
>
> LSE of β achieves minimum MSE among all **linear unbiased estimators**
>
> but this does not mean that $\hat{\beta}^{LSE}$ is the best estimator among all class

4. Test statistics from Linear Regression

- T statistic -> 유도

> $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$
> $T = \frac{\hat{\beta}_j^{LSE} - \beta_j}{\sqrt{\sigma^2 C_{jj}}} \sim t(n-p-1)$ where C_{jj} is jth diagonal element of $(X^TX)^{-1}$
> reject H_0 if $T > t_{0.975,n-p-1}$ or $T < t_{0.025,n-p-1}$

- F statistic -> 유도

> $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs $H_1 : notH_0$
> $F = \frac{V_1/k1}{V_2/k_2} \sim F(k1, k2)$, where $V_1 = SSE_{reduced} - SSE_{full}, \quad V_2 = SSE_{full}, k_1 and\ k_2$ are corresponding df
>
> reject H_0 if $F > F_{0.05}$

5. Multiple Regression from Simple Univariate Regression

- When all columns of Z is orthogonal, LSE fitting results are like below

> $\hat{\beta} = (Z^TZ)^{-1}Z^Ty = (z_0^Tz_0)^{-1}z_0^Ty+\cdots+(z_p^Tz_p)^{-1}z_p^Ty$
>
> which is same as take summation of Simple Univariate Regression results

- By using this, Multiple Regression can be fit by Succesive orthogonalization (called Gram-Schmidt precedure)

> $z_0 = \mathbf{x_0} = \mathbf{1}$
>
> $z_1 = \mathbf{x_1} - z_0\gamma_{01}$
>
> $\vdots$
>
> $z_p = \mathbf{x_p} - (\sum_{i=0}^{p} z_i\gamma_{ip})$, where $\gamma_{ij} = (z_i^Tz_i)^{-1}z_i^T\mathbf{x_j}$, single univariate regression coef on $\mathbf{x_j}$ by z_i
>
> because all z_i are orthogonal to each other, all $\hat{\beta}$ coef is fitted by univariate regression

- equations above can be expressed like below

> $\mathbf{x_0} = z_0$
>
> $\mathbf{x_1} = \gamma_{01}z_0 + z_1$
>
> $\vdots$
>
> $\mathbf{x_p} = \gamma_{0p}z_0 + \gamma_{1p}z_1 + \cdots + \gamma_{p-1\,p}z_{p-1} + z_p$

- these can be expressed as matrix form , which is called QR decomposition

> $$[\mathbf{x_0}\ \mathbf{x_1}\ \cdots \mathbf{x_p}] = [\mathbf{z_0}\ \mathbf{z_1}\ \cdots \mathbf{z_p}] \begin{bmatrix} 1 & \gamma_{01} & \gamma_{02} & \cdots & \gamma_{0p} \\ 0 & 1 & \gamma_{12} & \cdots & \gamma_{1p} \\ 0 & 0 & 1 & \cdots & \gamma_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$
>
> $$\begin{aligned} X &= Z\Gamma \\ &= ZD^{-1}D\Gamma \qquad \text{where } D = diag(\|\mathbf{z_0}\|, \cdots, \|\mathbf{z_p}\|) \\ &= QR \end{aligned}$$

- Backward fitting by QR decomposition needs lower computing power