

Multiple linear regression model with p parameters fit by least squares (LS) on $\{(x_1, y_1), \dots, (x_N, y_N)\} = \mathcal{T}$
 \uparrow drawn at random from a population $p(x, y)$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{LS estimate}$$

\downarrow
 $p \times N \quad N \times 1$

Test data $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population \rightarrow if we use LS on test data we get

$$\tilde{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$$

\downarrow
 $p \times M \quad M \times 1$

Now: $R_{tr}(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \beta^T x_i)^2$ and

$$R_{te}(\beta) = \frac{1}{M} \sum_{j=1}^M (\tilde{y}_j - \beta^T \tilde{x}_j)^2$$

* PROVE THAT $E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})]$
 \uparrow taken over $p(x, y)$

$$\text{THUS: } E[R_{tr}(\hat{\beta})] = E\left[\frac{1}{N} \sum_{i=1}^N (y_i - \hat{\beta}^T x_i)^2\right]$$

$$E[R_{te}(\hat{\beta})] = E\left[\frac{1}{M} \sum_{j=1}^M (\tilde{y}_j - \hat{\beta}^T \tilde{x}_j)^2\right]$$

OBSERVE: $\tilde{\beta}$ minimizes $R_{te}(\beta)$

FIRST: we assume that $M=N$, start with $R_{te}(\hat{\beta})$

$$\underbrace{\frac{1}{N} \sum_{j=1}^N (\tilde{y}_j - \hat{\beta}^T \tilde{x}_j)^2}_{R_{te}(\hat{\beta})} \geq \underbrace{\frac{1}{N} \sum_{j=1}^N (\tilde{y}_j - \tilde{\beta}^T \tilde{x}_j)^2}_{R_{te}(\tilde{\beta})}$$

Since $\tilde{\beta}$ minimize $R_{te}(\beta)$, so
 $R_{te}(\hat{\beta}) \geq R_{te}(\tilde{\beta})$

Now take E on both sides

$$E(R_{te}(\hat{\beta})) \geq E(R_{te}(\tilde{\beta}))$$

||

$$E(R_{tr}(\hat{\beta}))$$

this works as a training set
with $\hat{\beta}$ as LS estimator

$$E(R_{te}(\hat{\beta})) \geq E(R_{tr}(\hat{\beta}))$$

Now, remains to discuss what happens when $M \neq N$.

$$E(R_{te}(\hat{\beta})) = E\left(\frac{1}{M} \sum_{j=1}^M (\tilde{y}_j - \hat{\beta}^T \tilde{x}_j)^2\right) = \frac{1}{M} \sum_{j=1}^M E\left[(\tilde{y}_j - \hat{\beta}^T \tilde{x}_j)^2\right]$$

↑
not dependent on \tilde{x}_j
since made from training data

$$= E((\tilde{y}_1 - \hat{\beta}^T \tilde{x}_1)^2) \quad \text{for example } j=1$$

$$= E\left(\frac{1}{N} \sum_{j=1}^N (\tilde{y}_j - \hat{\beta}^T \tilde{x}_j)^2\right)$$

so

$$E(R_{te}(\hat{\beta})) = E(R_{tr}(\hat{\beta})) \geq E(R_{tr}(\hat{\beta}))$$

↑ ↑
with M with N

All good!