ELS Ex 2.9 page 40

Multiple linear regression model with p parameter fit by least squares (LS) on $\{(x_1, y_1), \ldots, (x_N, y_N)\} = \mathcal{T}$

$\uparrow$ drawn at random from a population $p(x,y)$

$\hat{\beta} = (X^T X)^{-1} \underbrace{X^T}_{p \times N} \underbrace{Y}_{N \times 1}$ LS estimate

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

$$\tilde{\beta} = (\tilde{X}^T \tilde{X})^{-1} \underbrace{\tilde{X}^T}_{p \times M} \underbrace{\tilde{Y}}_{M \times 1}$$

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

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

⊛ PROVE THAT $E\left[R_{tr}(\hat{\beta})\right] \leq E\left[R_{te}(\hat{\beta})\right]$

taken over $p(x,y)$

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

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

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

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

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

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

$$\| \\ E\left(R_{tr}(\hat\beta)\right)$$

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

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

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

not dependent on $\tilde x_j$ since made from training data

$$= E\left((\tilde y_1 - \hat\beta^T \tilde x_1)^2\right)$$ 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

$$E\left(\underset{\text{with } M}{R_{te}}(\hat\beta)\right) = E\left(\underset{\text{with } N}{R_e}(\hat\beta)\right) \geq E\left(R_{tr}(\hat\beta)\right)$$

All good!