Statistical Learning - Vienna, Fall2017

Homework problem 1

Problem ESL 2.9

Consider a linear regression model, fit by least squares to a set of training examples $T = \{(X_1, Y_1), ..., (X_N, Y_N)\}$, drawn i.i.d from some population. Let $\hat{\beta}$ be the linear least squares estimate. Suppose we also have some other ("test") data drawn independently from the same distribution $\{(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_M, \tilde{Y}_M)\}$. Prove that:

$$\frac{1}{N}\mathbb{E}(\sum_{i=1}^{N}(Y_i - X_i^T\hat{\beta})^2) \leq \frac{1}{M}\mathbb{E}(\sum_{i=1}^{M}(\tilde{Y}_i - \tilde{X}_i^T\hat{\beta})^2),$$

that is, the expected squared error for training is always smaller than for test in least squares fitting. Note that the values X are also random variables here, and the expectation is over everything that is random, including $X, Y, \tilde{X}, \tilde{Y}$ and $\hat{\beta}$.

Hint: There are several ways to prove this. One starts from considering the best possible linear model:

$$\beta^* = (E(XX^T))^{-1}E(XY),$$

and comparing both sides to it.

Note: Students who find more than one valid way to prove the result will get a bonus grade.