To Prove :

For two regression models f and f' trained on different subsets of training dataset, the distance measure $[f(x) - f'(x)]^2$ and the true squared error $[f(x) - y]^2$ (when the prediction for y is $f(x)$) are monotonically increasing under expectation.

Proof for Linear Models :

Consider a linear model, $\mathbf{y} = \mathbf{X}\beta + \epsilon$
where, $\mathbf{X}_{n\times m}$ is the data matrix with n independent observations and m covariates.
$\beta \in R^m$ is the parameter vector
$\mathbf{y} \in R^n$ is the response vector and
$\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ where ϵ_i are independent random variables with zero mean and variance of σ_y^2 for $i = 1(1)n$.

Assume we have centered covariates such that $x_{i,1} = 1$ and $E[x_{i,j}] = 0$ for all $j \neq 1$, where $j = 1$ corresponds to the intercept. Also we assume that the axes of covariates have been rotated such that the covariates are uncorrelated. $E[x_{i,j}x_{i,k}] = \sigma^2_{j,k}\delta_{j,k}$, where $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise

If $C \subseteq \{1, 2, ..., m\}$, Let X^C denote the matrix X with covariates corresponding to set C, i.e. if $C = \{c_1, c_2, ..., c_l\}$ then $X^C = [x^{(c_1)}, x^{(c_2)}, ..., x^{(c_l)}]$, where $x^{(i)}$ is the i'th column of X.

Also, If $S \subseteq \{1, 2, ..., n\}$, Let X_S denote the matrix X with observations corresponding to set S, i.e. if $S = \{s_1, s_2, ..., s_l\}$ then $X_S = [x_{s_1}, x_{s_2}, ..., x_{s_l}]^T$, where x_i is the i'th row of X.

If C_1, C_2 be two subsets of $\{1, 2, ..., m\}$, such that $|C_1| = m_1, |C_2| = m_2$; and S_1, S_2 be two subsets of $\{1, 2, ..., n\}$, such that $|S_1| = n_1, |S_2| = n_2$.

For an observation $x = (x_1, x_2, ..., x_m)^T$, let the prediction $\hat{y}$ from the two OLS models be as follows :

- The model is trained on data $X^{C_1}_{S_1}$ and the response y_{S_1} corresponding to observations S_1. So, $\hat{y} = f(x) = x^T\hat{\beta}_1$. $\hat{\beta}_1$ is the corresponding parameter vector for all covariates i.e. 0 for all covariates not in S_1 and the OLS values for all covariates in S_1.
- The model is trained on data $X^{C_2}_{S_2}$ and the response y_{S_2} corresponding to observations S_2. So, $\hat{y} = f(x) = x^T\hat{\beta}_2$. $\hat{\beta}_2$ is the corresponding parameter vector for all covariates i.e. 0 for all covariates not in S_2 and the OLS values for all covariates in S_2.

For the first model,

Then the MSE $E[(f(x) - y)^2]$ is monotonically related with the expected squared distance measure between f and f'.

$$E[(f(x) - y)^2] = (1 + n/n')^{-1}E[(f(x) - f'(x))^2] + \sigma_y^2$$

Proof

For a dataset of size n, the OLS estimate of β, denoted by $\hat{\beta}$, is a random variable that obeys a normal distribution with a mean of β and a covariance given by $n^{-1}\Sigma$

$$Var(\hat{\beta}) = \sigma_y^2(X^TX)^{-1} = \sigma_y^2(n\ diag\{1, \sigma_{2,2}^2, ..., \sigma_{m,m}^2\})^{-1} = n^{-1}\Sigma$$

where $\Sigma = \sigma_y^2 diag\{1, \sigma_{2,2}^{-2}, ..., \sigma_{m,m}^{-2}\}$.

Hence,

$$E[(f(x) - y)^2] = Var(x^T\hat{\beta} - y) = x^T(n^{-1}\Sigma)x + \sigma_y^2$$

and

$$E[(f(x) - f'(x))^2] = Var(x^T\hat{\beta} - x'^T\hat{\beta}') = x^T(n^{-1}\Sigma)x + x'^T(n'^{-1}\Sigma')x' = x^T(n^{-1} + n'^{-1}\Sigma)x$$

Replacing $x^T\Sigma x$ in first equation we get ,

$$E[(f(x) - y)^2] = (1 + n/n')^{-1}E[(f(x) - f'(x))^2] + \sigma_y^2$$

Hence proved.