Ex 2.9

W.t.s. $E[R_{tr}(\hat{\beta}_{tr})] \leq E[R_{te}(\hat{\beta}_{tr})]$

Note: $\hat{\beta}_{tr}$ refers to the least sq. estimates from training data $(x_1, y_1) \ldots (x_N, y_N)$. Similarly, $\hat{\beta}_{te}$ refers to the least sq. estimates from testing data $(\tilde{x}_1, \tilde{y}_1) \ldots (\tilde{x}_M, \tilde{y}_M)$.

$$R_{te}(\hat{\beta}_{tr}) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}_{tr}^T \tilde{x}_i)^2 \geq \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2$$

(as $\hat{\beta}_{te}$ is defined as the $\hat{\beta}$ that minimises this eqn.)

So, $E(R_{te}(\hat{\beta}_{tr})) \geq E(1/M \sum_i (\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2$

$= E[(\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2]$ (since any arbitrary $(\tilde{x}_i, \tilde{y}_i)$ pair is drawn from the same population)

$= E[1/N \sum_{i=1}^{N} (\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2]$ (same logic)

Since N test points are drawn from the same population as N train points, this is equivalent to:

$= E[1/N \sum_{i=1}^{N} (y_i - \hat{\beta}_{tr}^T x_i)^2] = E[R_{tr}(\hat{\beta}_{tr})]$

$\Rightarrow E[R_{te}(\hat{\beta}_{tr})] \geq E[R_{te}(\hat{\beta}_{tr})]$