2.7 (a)

linear regression:

$$\hat{Y} = X\hat{\beta} \quad \text{where} \quad \hat{\beta} = (X^T X) X^T y$$

$$\Rightarrow \hat{f}(x_0) = x_0^T \hat{\beta}$$

$$= \underbrace{x_0^T (X^T X) X^T}_{} \underbrace{y}_{(n \times 1)}$$

$$\text{dimensionality} = \underbrace{(1 \times P)(P \times n)(n \times P)(P \times n)}_{(1 \times n)}$$

Rewrite as: $[a_1 \, a_2 \dots a_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = Ay$

$$\Rightarrow \hat{f}(x_0) = \sum_{i=1}^{n} a_i y_i$$

and since A is a function of x_0, X we can express this as: $a_i = l_i(x_0 : X)$

K-nearest-neighbours:

$$\hat{f}(x_0) = 1/k \sum_{x_i \in N_k(x_0)} y_i$$

$$= \sum_{x_i \in N_k(x_0)} (1/k) y_i$$

$$= \sum_{i=1}^{n} \delta_i (1/k) y_i \quad \text{where} \quad \delta_i = \begin{cases} 1 & \text{if } x_i \in N_k(x_0) \\ 0 & \text{else} \end{cases}$$

$$= \sum_{i=1}^{n} l_i(x_0 : X) y_i$$

2.7 (b) + (c) + (d)

$$Var(X) = E(X^2) - E(X)^2$$
$$Bias_\theta(\hat{\theta}) = E_{x|\theta}(\hat{\theta}) - \theta$$

Part (b):

Now, $E_{y|x}(f(x_0) - \hat{f}(x_0))^2$

$$= E_{y|x}(f(x_0)^2 - 2f(x_0)\hat{f}(x_0) + \hat{f}(x_0)^2)$$

$$= f(x_0)^2 - 2f(x_0)E_{y|x}(\hat{f}(x_0)) + E_{y|x}(\hat{f}(x_0)^2)$$

as $f(x_0)$ has no dependance on the training data

$$= (f(x_0) - E_{y|x}(\hat{f}(x_0)))^2 - E_{y|x}(\hat{f}(x_0))^2 + E_{y|x}(\hat{f}(x_0)^2)$$

$$= \left[Bias_{f(x_0)}(\hat{f}(x_0))\right]^2 + Var(\hat{f}(x_0))$$

Part (c) follows the exact same steps as (b) and also finds:

$$E_{x,y}(f(x_0) - \hat{f}(x_0))^2$$

$$= \left[Bias_{f(x_0)}(\hat{f}(x_0))\right]^2 + Var(\hat{f}(x_0))$$

(d) The bias measures the systematic error - how much the expected model predictions differ from the true values.

The variance, however, is simply a measure of the variation of the expected model predictions.