Problem 4

1. $\mathbb{E}(Y) = \text{argmin}_c \; \mathbb{E}[(Y-c)^2]$

$$= \text{argmin}_c \; \mathbb{E}[Y-c]^2$$

with the term $[Y-c]^2$ being quadratic (convex), the value of c that would minimize the term would be $\mathbb{E}(Y)$ since the minimum is 0.
In order to get to the lowest possible error in linear regression using an estimator (c), it needs to be equal to $\mathbb{E}(Y)$ which in turn is equal to $f(X)$ which would be the best case scenario. That way we achieve 0 reducible error & only the irreducible error remains.

2. $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

- $R^2 \stackrel{?}{=} Cor(X, Y)^2$

$RSS = \sum_i^n (y_i - \hat{y}_i)^2$; ④ $TSS = \sum_i^n (y_i - \bar{y})^2$

① $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$; ② $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (in linear regression)

③ $\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$ (〃)

$$
\begin{aligned}
RSS &= \sum_i^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \quad \text{using ① \& ②} \\
&= \sum_i^n (\hat{\beta}_1 (\bar{x} - x_i) - (\bar{y} - y_i))^2 \\
&= \hat{\beta}_1^2 \sum_i^n (\bar{x} - x_i)^2 - 2\hat{\beta}_1 \sum_i^n (\bar{x} - x_i)(\bar{y} - y_i) + \sum_i^n (\bar{y} - y_i)^2 \\
&= \hat{\beta}_1 (\hat{\beta}_1 \sum_i^n (\bar{x} - x_i)^2 - 2\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i)) + \sum_i^n (\bar{y} - y_i)^2 \quad \text{use ③ here} \\
&= \hat{\beta}_1 (\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i) - 2\sum (\bar{x} - x_i)(\bar{y} - y_i)) + \sum_i^n (\bar{y} - y_i)^2 \\
&= \sum_i^n (\bar{y} - y_i)^2 - \hat{\beta}_1 \sum_i^n (\bar{x} - x_i)(\bar{y} - y_i) \quad \text{use ③} \\
&= \sum_i^n (\bar{y} - y_i)^2 - \frac{(\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i))^2}{\sum_i^n (\bar{x} - x_i)^2} \\
&= \sum_i^n (\bar{y} - y_i)^2 \left(1 - \frac{(\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i))^2}{\sum_i^n (\bar{x} - x_i)^2 \sum_i^n (\bar{y} - y_i)^2}\right) \quad ⑤
\end{aligned}
$$

$R^2 = 1 - \frac{RSS}{TSS}$ using ④ & ⑤ $= \frac{(\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i))^2}{\sum_i^n (\bar{x} - x_i)^2 \sum_i^n (\bar{y} - y_i)^2}$ ⑥

we have $Cor(X, Y) = \frac{\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i)}{\sqrt{\sum_i^n (\bar{x} - x_i)^2} \sqrt{\sum_i^n (\bar{y} - y_i)^2}}$ ⑦

from squaring ⑦ we get ⑥ $\rightarrow R^2 = Cor(X, Y)^2$

- $R^2 = Cor(Y, \hat{Y})^2$