Q1)

Part a)

$$f^* = \arg\min_f E[(f(X) - Y)^2] = \arg\min_f E[f(X)^2 - 2f(X)Y + Y^2]$$
$$= \arg\min_f E[f(X)^2] - 2E[f(X)Y] + E[Y^2] = K$$

$$\frac{\partial K}{\partial f(X)} = 0 = 2f(X) - 2E[Y] \rightarrow f^*(x) = E[y]$$

Part b)

$$P(Y = y_k|X) \propto \exp\left(w_{k_0} + \sum_1^d w_{k_i} X_i\right) = \exp(w_{k_0}) \exp\left(\sum_1^d w_{k_i} X_i\right)$$

This formula expresses SoftMax model $softmax(z) = \frac{e^z}{\sum_1^d e^{z'}}$, and it is from exponential family. Also, $P(Y = y_k|X)$ is a GLM.

Part c)

$\hat{y} = P(Y = y_k|X)$ selects the class with most probability.

Q2)

Part a)

In order to create less complex (parsimonious) model when you have a large number of features in your dataset, some of the Regularization techniques used to address over-fitting and feature selection are:

1. L1 Regularization (Lasso Regression)
2. L2 Regularization (Ridge Regression)

L1 adds "*absolute value of magnitude*" of coefficient as penalty term to the loss function.

$$\text{Cost Function: } \sum_1^n \left(y_i - \sum_1^p x_{ij}\beta_j\right)^2 + \lambda \sum_1^p |\beta_j|$$

It shrinks the less important feature's coefficient to zero.

L2 adds "squared magnitude" of coefficient as penalty term to the loss function.

$$\text{Cost Function: } \sum_1^n \left(y_i - \sum_1^p x_{ij}\beta_j\right)^2 + \lambda \sum_1^p \beta_j^2$$

The difference between these techniques is that L2, unlike L1, does not reduce the coefficients to zero. Also, by using L2 we will have a more complicated model in comparison to L1.

Part b)

$J(w) = \left||Xw - Y|\right|_2^2 + \lambda\left||w|\right|_2^2 = (Xw - Y)^T(Xw - Y) + \lambda w^T w$

$= (Xw)^T Xw - (Xw)^T Y - Y^T(Xw) + Y^T Y + \lambda w^T w$

$= w^T(X^T X)w - 2Y^T(Xw) + Y^T Y + \lambda w^T w$

$\dfrac{dJ(w)}{dw} = (X^T X)w - X^T Y + \lambda w^T$

Initial guess: ω_0: $J(\omega_0) = ||X\omega_0 - Y||_2^2 + \lambda||\omega_0||_2^2 = (X\omega_0 - Y)^T(X\omega_0 - Y) + \lambda{\omega_0}^T\omega_0$

Second xxx: $\omega_1 = \omega_0 - \frac{J(\omega_0)}{J'(\omega_0)} = \omega_0 - J(\omega_0)\big(J'(\omega_0)\big)^{-1}$

$= \omega_0 - [{\omega_0}^T(X^T X)\omega_0 - 2Y^T(X\omega_0) + Y^T Y + \lambda{\omega_0}^T\omega_0][(X^T X)\omega_0 - X^T Y + \lambda{\omega_0}^T]^{-1}$

$= \omega_0 - [{\omega_0}^T(X^T X)\omega_0 - 2Y^T(X\omega_0) + Y^T Y + \lambda{\omega_0}^T\omega_0][\dots\dots]$

$= \omega_0 - [{\omega_0}^T(X^T X)\omega_0(\dots) - 2Y^T(X\omega_0)(\dots) + Y^T Y(\dots) + \lambda{\omega_0}^T\omega_0(\dots)]$

$= X^T X\omega_0 - \lambda\omega_0 = X^T Y$

$\rightarrow \omega_1 = w^* = (X^T X - \lambda I)^{-1} X^T y$

Q3)

n	Σx	Σy	(Σxy)	(Σx^2)	$(\Sigma x)^2$
10	189	561	12521	4173	35721

Part a)

$y = \beta_0 - \varepsilon_i$

$$\rightarrow \beta_0 = (X^T X)^{-1} X^T y = \frac{1}{10} X^T y = \bar{y} = \frac{561}{10} = 56.1$$

Part b)

$y = \beta_1 x - \varepsilon_i = 3.192x + 4.2297$

$$\rightarrow \beta_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = 3.1920$$

Part c)

The second one shows the real equation with its natural noise that can be reduced but cannot be removed and used to find y's range. The first one is estimation of the second one and obtained y using this equation could have some errors.

Part d)

$\hat{y} = 25 - 0.5x \;\rightarrow\; \hat{y}(6) = 25 - 3 = 22$

Due to natural noise, it cannot be said that 22 is the exact number.

$y_{ture} = 22 + \epsilon$

Part e)

$n = 16$ and $SSE = 7$

$$\sigma^2 = MSE = \frac{SSE}{n-2} = \frac{7}{16-2} = 0.5$$