Q 1)

Update rule using Sigmoid activation function.

$$z_1 = w_1 x + b_1$$
$$a_1 = \text{sigmoid}(z_1)$$
$$z_2 = w_2 a_1 + b_2$$
$$a_2 = \text{linear}(z_2) \quad \{\text{here linear is used as we need regression}\}$$

$$\hat{y} = a_2 = g(z_2)$$

$$\text{MSE (loss function)} = \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2$$

$$\frac{dL}{da_2} = \frac{d}{da_2}(y^2 + a_2^2 - 2y a_2)$$
$$= (0 + 2a_2 - 2y)$$
$$= 2(a_2 - y)$$

Applying chain rule,

$$\frac{dL}{dw_2} = \frac{dL}{da_2} \times \frac{da_2}{dz_2} \times \frac{dz_2}{dw_2}$$
$$= 2(a_2 - y) \times 1 \times (a_1 + 0)$$

$$\boxed{\frac{dL}{dw_2} = 2(a_2 - y) a_1}$$

* $\frac{dL}{db_2} = \frac{dL}{da_2} \times \frac{da_2}{dz_2} \times \frac{dz_2}{db_2}$

$= 2(a_2 - y) \times \frac{dz_2}{dz_2} \times \frac{d}{db_2}(w_2 a_1 + b_2)$

$\frac{dL}{db_2} = 2(a_2 - y)$

* $\frac{dL}{dw_1} = \frac{dL}{da_2} \cdot \frac{da_2}{dz_2} \cdot \frac{dz_2}{da_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz_1}{dw_1}$

$= 2(a_2 - y) \cdot (1 - a_1) a_1 \cdot w_2 \cdot x$

* $\frac{dL}{db_1} = \frac{dL}{da_2} \cdot \frac{da_2}{dz_2} \frac{dz_2}{da_1} \frac{da_1}{dz_1} \frac{dz_1}{db_1}$

$\frac{dL}{db_1} = 2(a_2 - y)(1 - a_1) a_1 \cdot w_2$

Comparision between MSE & binary classification using log loss

MSE updates for regression	Updates for binary classification using logloss.
$\frac{dL}{db_1} = 2(a_2-y)\, w_2(1-a_1)\, a_1$	$\frac{dL}{db_1} = (a_2-y)\, w_2 a_1(1-a_1)$
$\frac{dL}{dw_1} = 2(a_2-y)\, w_2(1-a_1)\, a_1 . x$	$\frac{dL}{dw_1} = (a_2-y)\, w_2 . a_1(1-a_1) . x .$
$\frac{dL}{db_2} = 2(a_2-y)$	$\frac{dL}{db_2} = 2(a_2-y)$
$\frac{dL}{dw_2} = 2(a_2-y) a_1$	$\frac{dL}{dw_2} = 2(a_2-y) a_1$

Updates for regression are two times the updates for classification.

Also, in regression we have linear function for output layer and sigmoid for an activation function, which is not the case in classification.