Multiple Linear Regression Equations

In multiple linear regression, we have several input features $x_1, x_2, \ldots, x_n$, and the goal is to fit a linear model to predict the output y.

1. Multiple Linear Regression Equation

The prediction for the output $\hat{y}^{(i)}$ for the i-th data point, based on multiple features, is given by:

$$\hat{y}^{(i)} = w_1 \cdot x_1^{(i)} + w_2 \cdot x_2^{(i)} + \cdots + w_n \cdot x_n^{(i)} + b$$

This can be written more compactly using vector notation:

$$\hat{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + b$$

Where:

- $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ is the vector of weights (parameters),
- $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]^T$ is the vector of features for the i-th data point,
- b is the bias (intercept term).

2. Cost Function

The cost function $J(\mathbf{w}, b)$ for multiple linear regression, typically the Mean Squared Error (MSE), is defined as:

$$J(\mathbf{w}, b) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)^2$$

Where:

- m is the number of data points,
- $y^{(i)}$ is the actual value for the i-th data point,
- $\hat{y}^{(i)}$ is the predicted value for the i-th data point.

3. General Gradient Descent Equation

The general gradient descent equation used to minimize the cost function $J(\theta)$ is:

$$\theta := \theta - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta}$$

Where:

- θ represents the parameters (e.g., $\mathbf{w}$ or b),
- α is the learning rate,
- $\frac{\partial J(\theta)}{\partial \theta}$ is the gradient of the cost function with respect to θ.

4. Gradients with Respect to the Parameters w and b

The partial derivatives (gradients) of the cost function with respect to each parameter in the vector $\mathbf{w}$ and the bias b are:

$$\frac{\partial J(\mathbf{w}, b)}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \quad \text{for each } j = 1, 2, \ldots, n$$

$$\frac{\partial J(\mathbf{w}, b)}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)$$

5. Update Equations for the Parameters w and b

Using gradient descent, the parameters are updated iteratively. The update rules for the weights $\mathbf{w}$ and bias b are:

$$w_j := w_j - \alpha \cdot \frac{\partial J(\mathbf{w}, b)}{\partial w_j} \quad \text{for each } j = 1, 2, \ldots, n$$

$$b := b - \alpha \cdot \frac{\partial J(\mathbf{w}, b)}{\partial b}$$

Substituting the gradients:

$$w_j := w_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \quad \text{for each } j = 1, 2, \ldots, n$$

$$b := b - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)$$

These equations iteratively adjust the parameters $\mathbf{w}$ and b to minimize the cost function, leading to the optimal values for the weights and bias.